SAC-Copula: Quality-Preserving Watermarking for Diffusion Language Models via Smooth Correlated Gumbel Fields

Baixin Li and Haiyun He
The Hong Kong University of Science and Technology (Guangzhou)
baixinli31@connect.hku.hk, haiyunhe@hkust-gz.edu.cn

Abstract

Watermarking diffusion language models (DLMs) requires mechanisms compatible with iterative parallel unmasking rather than autoregressive decoding. Existing sampling-based watermarking methods typically inject position-wise i.i.d. perturbations, which can be poorly aligned with DLM decoding dynamics and degrade generation quality. We propose SAC-Copula, a quality-preserving watermarking method for DLMs based on smooth, locally correlated Gumbel perturbation fields constructed via a Gaussian copula. We further develop a SAC-aware detector using covariance-aware filtering and native-sample calibration. Mechanism-level analysis shows that local correlation reduces latent perturbation roughness and better matches iterative refinement dynamics. Experiments on LLaDA show that SAC-Copula achieves a favorable quality–detectability trade-off compared with existing baselines. In particular, further evaluations on Dream-7B and additional datasets show that SAC-Copula substantially improves PPL tail stability over the i.i.d. Gumbel baseline, while maintaining strong low-FPR detectability and competitive overall generation quality. Additional token-edit stress tests further assess watermark robustness under controlled synchronization drift. Code is available at https://github.com/PunkyKnife/SAC-Copula.

1 Introduction

As Large Language Models (LLMs) become widely deployed, it becomes increasingly critical to establish a safety line between LLMs and humans, where one important brick is to distinguish the model-generated content from the human-created one. Watermarking offers a practical route: a generator embeds a private statistical signal during content decoding, and a detector later tests whether a text carries that signal via a shared secret key.

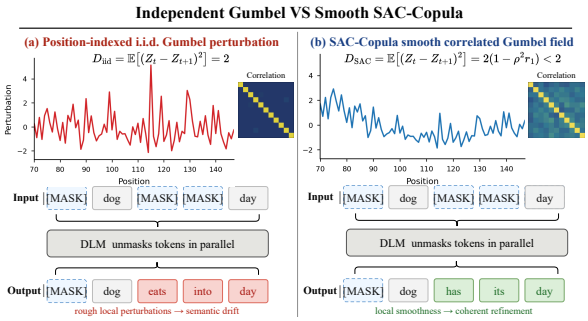


Figure 1: Motivation for smooth watermark fields in DLM decoding. Position-indexed i.i.d. Gumbel perturbations create independent high-frequency noises across neighboring positions, whereas SAC-Copula introduces controlled low-lag dependence while preserving one-coordinate Gumbel marginals. Token examples are illustrative rather than deterministic generation outcomes.

Most existing language-model watermarking methods are developed for autoregressive LLMs, including green-list token biasing, sampling-based or distortion-free schemes, robust watermarking variants, and multi-bit identification mechanisms (Kirchenbauer et al., 2023; Kuditipudi et al., 2024; Zhao et al., 2024; Yoo et al., 2024). In these methods, watermark signals are typically synchronized with the left context and tokens are generated sequentially from left to right conditioned on the prefix and watermark signal. In contrast, Diffusion Language Models (DLMs), such as LLaDA (Nie et al., 2025), generate text through iterative masked denoising, where multiple positions are updated in parallel and previously generated tokens can be revised across steps. This non-sequential and dynamically coupled decoding process makes watermark design fundamentally different from the autoregressive setting. In particular, watermarking schemes based on i.i.d. injected signals are poorly matched to DLM decoding dynamics, which can disrupt local refinement consistency across denoising steps. This may lead to degraded generation quality, including unstable neighboring predictions, semantic

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 20961–20984
October 24-29, 2026 ©2026 Association for Computational Linguistics

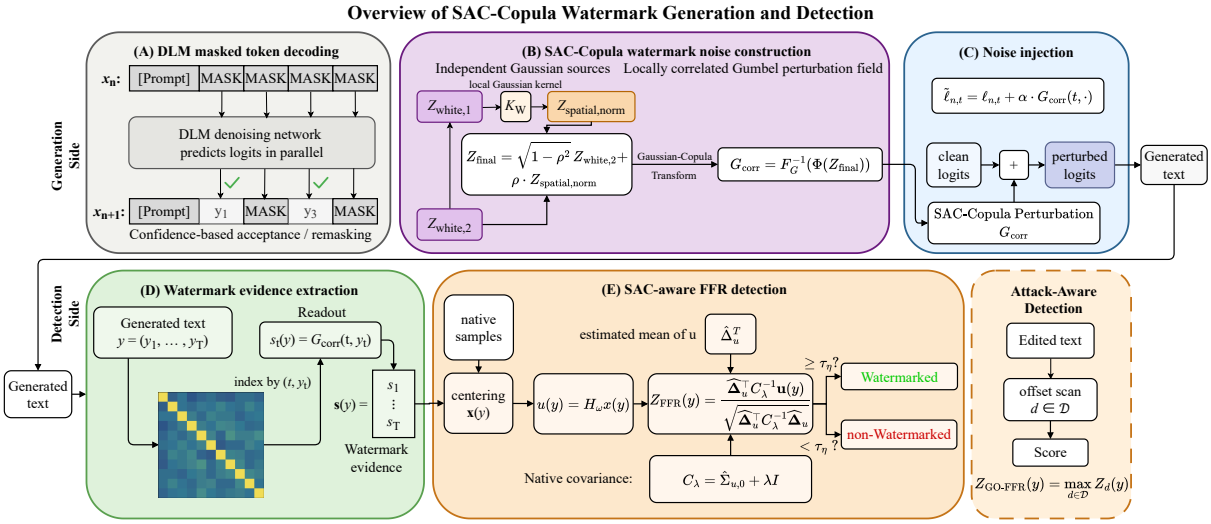


Figure 2: Overview of SAC-Copula. The generation side constructs a locally correlated Gumbel tape through variance-normalized convolution, dual-stream mixing, and a Gaussian-copula transform, then injects it at currently masked generated positions during DLM refinement. The detection side reads token-level evidence and applies SAC-aware filtering with thresholds calibrated on native, unwatermarked samples. GO-FFR is shown only as an auxiliary diagnostic for controlled insertion/deletion edits.

inconsistency, degraded fluency, and refinement oscillation. Figure 1 shows an illustrative example that compares i.i.d. and correlated perturbations.

1.1 Related Works

Autoregressive and sampling-based watermarking. A large body of LLM watermarking work embeds keyed statistical signals during autoregressive decoding and detects them under a false-positive constraint (Kirchenbauer et al., 2023; Li et al., 2025). Stronger watermark signals can improve detection power but may degrade fluency, semantic fidelity, or diversity, motivating selective, adaptive, and quality-aware objectives (Wang et al., 2025; Giboulot and Furon, 2024; He et al., 2025). A related line uses Gumbel-max, inverse-transform, or distortion-free sampling ideas to embed keyed randomness under specific sampling and security assumptions (Aaronson, 2023; Fairoze et al., 2023; Christ et al., 2024). This paper follows the sampling-based watermarking paradigm, while explicitly adapting the design to DLM decoding dynamics.

DLM watermarking. Some recent works have started to explore designing watermarking methods adapted to DLMs. Gloaguen et al. (2026) propose the first watermark tailored for DLMs by applying the green-red list watermarking (Kirchenbauer et al., 2023) in expectation over partially observed contexts and promoting tokens that strengthen future watermark signals. To address the non-sequential decoding process, some works (Wu et al., 2025; Raban et al., 2026) further consider synchronizing watermark signals using both left and right contexts. Instead of directly perturbing token probabilities, Hong and No (2026) embed watermark by steering the decoding order to improve quality preservation in DLMs.

More closely related to our work, Bagchi et al. (2026) apply Gumbel-max watermarking (Aaronson, 2023) to diffusion decoding using sequence-position-indexed randomness. Although this approach is theoretically distribution-preserving, the injected i.i.d. Gumbel perturbations are not well aligned with DLM decoding dynamics, as discussed earlier, and can degrade generated text quality in practice. Existing Gumbel-based methods primarily focus on how randomness is indexed across positions. In contrast, we argue that the *joint structure* of the perturbation field is itself a critical design dimension for DLM watermarking. Under iterative refinement decoding, independent high-frequency perturbations across neighboring positions can conflict with local consistency. Motivated by this observation, we shift the design focus from injecting independent randomness to designing structured perturbation fields whose geometry and smoothness better align with diffusion decoding dynamics and detector evidence geometry.

The main insight of our paper is that locally smooth and correlated watermark perturbations

may be more compatible with DLM decoding than high-frequency i.i.d. perturbations, motivating the hypothesis of improved generation stability. Based on this insight, we propose **SAC-Copula**, which constructs correlated Gumbel perturbation fields via a Gaussian copula (cf. Figure 1). Under suitable assumptions, SAC-Copula preserves Gumbel marginals while introducing controlled local dependence across neighboring token positions. This structured perturbation field is designed to better align with DLM refinement dynamics and supports a strong quality–detectability trade-off.

1.2 Main Contributions

Our contributions can be summarized as follows:

- We introduce **SAC-Copula**, a DLM watermarking method with **S**mooth and **A**uto-**C**orrelated Gumbel Copula perturbations through a Gaussian **Copula**, which preserves one-coordinate Gumbel marginals under mild assumptions. This is a DLM-aware watermarking perspective that treats the joint geometry of the perturbation field as a key design dimension for DLM decoding.
- We provide a mechanism-level analysis showing that positive low-lag dependence reduces latent perturbation roughness, decreasing the local-difference energy from $D_{\mathrm{i.i.d}} = 2$ to $D_{\mathrm{SAC}} = 2(1 - \rho^2 r_1)$ and providing a mechanistic rationale for the DLM-refinement hypothesis.
- We develop a SAC-aware detector that reads the induced local evidence geometry through covariance-aware filtering and native-sample calibration. For controlled edit diagnostics, we further include a global-offset scan that provides coarse synchronization recovery under token–signal misalignment.
- Experiments on LLaDA show that SAC-Copula achieves a strong quality–detectability trade-off, with strong clean detectability and a balanced multi-dimensional quality profile. Relative to the matched i.i.d. Gumbel baseline, the clearest generation-side gain is substantially improved upper-tail stability. We further report mild token-edit stress tests as diagnostic analyses of residual watermark evidence.

2 Watermark Injection and Detection

SAC-Copula replaces position-indexed i.i.d. Gumbel perturbations with a smooth, locally correlated Gumbel field for DLM decoding. The generation side defines this field, while the detector provides a matched readout for the induced evidence geometry. Figure 2 summarizes the pipeline.

Baseline i.i.d. Gumbel Watermarking. Let $\mathcal{V}$ be the token vocabulary, $j \in \mathcal{V}$ a token, $t \in [T]$ a generated-token position, and n a DLM denoising step. The direct baseline applies position-indexed i.i.d. Gumbel perturbations (Bagchi et al., 2026; Aaronson, 2023), $G_{\mathrm{i.i.d.}}(t, j) \sim \mathrm{Gumbel}(0, 1)$, to masked-position logits:

$$\widetilde{\ell}_{n,t,j} = \ell_{n,t,j} + \alpha G_{\mathrm{i.i.d.}}(t, j), \quad t \in [T], j \in \mathcal{V}.$$

Here $\ell_{n,t,j}$ is the clean logit and α is the perturbation scale. In the controlled mechanism comparison used in this paper, the direct i.i.d. Gumbel baseline is the $\rho = 0$ endpoint of the same tape family: setting $\rho = 0$ removes cross-position correlation while retaining standard-Gumbel candidate marginals and the same perturbation scale α. In the main comparison, both settings use $\alpha = 1.0$ and temperature 1.0, with the same key/tape interface, prompts, seeds, generation length, diffusion steps, masked-position injection, decoding, and remasking pipeline. We do not introduce or tune a separate δ-style logit-bias strength for this matched i.i.d.–SAC comparison; α is fixed at 1.0, while W, σ_K, and ρ parameterize the cross-position dependence structure.

2.1 SAC-Copula Perturbation Field

SAC-Copula constructs a locally correlated Gaussian latent field and maps each coordinate to a standard Gumbel variable through a Gaussian copula. It starts from two independent key-seeded standard Gaussian position–vocabulary tapes $Z_{\mathrm{white},1}, Z_{\mathrm{white},2}$, with $Z_{\mathrm{white},r}(t, j) \sim \mathcal{N}(0, 1)$. Let W be an odd window, $h = (W - 1)/2$, and σ_K the kernel shape parameter. For $k \in \{-h, \ldots, h\}$, define

$$K_W[k] = \frac{\exp\left(-k^2/(2\sigma_K^2)\right)}{\sum_{m=-h}^{h} \exp\left(-m^2/(2\sigma_K^2)\right)}.$$

For full-window positions, the normalized spatial branch is

$$\widetilde{Z}_{\mathrm{spatial}}(t, j) = \frac{\sum_{k=-h}^{h} K_W[k] Z_{\mathrm{white},1}(t - k, j)}{\sqrt{\sum_{k=-h}^{h} K_W[k]^2}}.$$

The final latent field mixes the independent white noise and the normalized spatial stream:

$$Z_{\mathrm{final}}(t, j) = \sqrt{1 - \rho^2}\, Z_{\mathrm{white},2}(t, j) + \rho\, \widetilde{Z}_{\mathrm{spatial}}(t, j),$$

where $\rho \in [0,1]$ controls the local dependence strength. Under the full-window or boundary-normalized convention, this normalization keeps each valid coordinate of Z_{final} standard normal. Appendices A.1 and A.2 state the boundary convention and prove the resulting one-coordinate marginal exactness.

The correlated Gumbel noise tape is obtained by the Gaussian-copula transform

$$G_{\text{corr}}(t,j) = F_G^{-1}(\Phi(Z_{\text{final}}(t,j))) \sim \text{Gumbel}(0,1),$$

where Φ is the standard Gaussian cumulative distribution function (CDF) and F_G is the standard Gumbel CDF. Appendix A.2 shows that this coordinate-wise transform preserves one-coordinate Gumbel marginals under assumptions stated in Appendix A.1, while introducing the joint noise tape structure. Since the perturbations are also independent across vocabulary items at each fixed position, it further implies that the single-position conditional sampling marginal is preserved:

$$P^{\text{SAC}}_{Y_t|\text{context}} = P^{\text{i.i.d.}}_{Y_t|\text{context}},$$

where Y_t denotes the token at the t-th position and the context denotes all information used to compute the current logits. This is a fixed-context, single-position statement, not invariance of the full generated-text joint distribution. Appendix A.3 derives the finite-window latent autocorrelation used by the detector design.

2.2 Injection into LLaDA Decoding

SAC-Copula noise is injected only into currently masked generated positions. At a masked generated position t, SAC-Copula perturbs logits before applying the decoding rule:

$$\widetilde{\ell}_{n,t,j} = \ell_{n,t,j} + \alpha G_{\text{corr}}(t,j), \quad t \in [T], j \in \mathcal{V},$$

Here $\ell_{n,t,j}$ is the contemporaneous clean logit from the current partially masked denoising state. The keyed tape G_{corr} is constructed independently of model logits; the logits determine how the fixed tape perturbation affects token selection, not how the tape itself is constructed. The perturbation scale α is an experimental configuration parameter. In the main configuration it matches the generation temperature, but this equality is not a theoretical requirement. Prompt tokens are not perturbed, and remasking confidence is computed from the unperturbed logits. The same token-position perturbation $G_{\text{corr}}(t,\cdot)$ is reused across denoising steps, indexed by generated token position and token but not by denoising step.

2.3 SAC-Aware Detector

For a clean fixed-alignment sequence $y = (y_1, \ldots, y_T)$, the detector reconstructs the same private keyed G_{corr} tape and indexes it with the observed token at each generated position, yielding token-level watermark evidence $\mathbf{s}(y) = (s_1(y), \ldots, s_T(y))^\top$, where

$$s_t(y) = G_{\text{corr}}(t, y_t).$$

The tape reconstruction depends on the private key rather than on a second model evaluation; detection does not recompute fully revealed logits or require an additional language-model forward pass. The legacy detector treats these values as exchangeable i.i.d. Gumbel evidence and applies an equal-weight z-score; Appendix B.1 gives the full old-score form. SAC-Copula intentionally induces low-lag evidence structure, so relying only on an i.i.d. equal-weight readout can discard useful local geometry. We therefore use the Full Filtered Ridge (FFR) detector, defined as a ridge-regularized matched linear statistic after SAC-aware local filtering and covariance calibration on native samples.

FFR first centers raw evidence using the Gumbel mean μ_G and a residual mean $\widehat{\boldsymbol{\mu}}_{x,0}$ estimated from held-out native (aka. unwatermarked) samples generated under the same decoding protocol:

$$\mathbf{x}(y) = \mathbf{s}(y) - \mu_G \mathbf{1} - \widehat{\boldsymbol{\mu}}_{x,0}.$$

It then applies a SAC-aware local feature map

$$\mathbf{u}(y) = H_\omega \mathbf{x}(y),$$
$$\omega_\tau \propto \exp\left(-\frac{\tau^2}{4\sigma_K^2}\right) \mathbf{1}\{|\tau| \le B\},$$

where H_ω is a filter motivated by the autocorrelation envelope of the SAC kernel (cf. Appendix B.2) and $B = \min(W-1, T-1)$. The lag template mirrors the Gaussian-envelope smoothing used by SAC-Copula, so H_ω aggregates evidence over the local low-lag structure introduced by the correlated perturbation field.

The filtered native-sample covariance and ridge covariance are

$$\widehat{\Sigma}_{u,0} = H_\omega \widehat{\Sigma}_{x,0} H_\omega^\top, \qquad C_\lambda = \widehat{\Sigma}_{u,0} + \lambda I.$$

Here $\widehat{\Sigma}_{x,0}$ is estimated from held-out native samples under the same detector configuration. Let $\widehat{\mathbf{\Delta}}_u$ be the filtered mean direction estimated from development or cross-fit data, not evaluation samples. The FFR score is

$$Z_{\mathrm{FFR}}(y) = \frac{\widehat{\mathbf{\Delta}}_u^\top C_\lambda^{-1} \mathbf{u}(y)}{\sqrt{\widehat{\mathbf{\Delta}}_u^\top C_\lambda^{-1} \widehat{\mathbf{\Delta}}_u}}.$$

The detector flags a sequence as watermarked when

$$Z_{\mathrm{FFR}}(y) \geq \tau_\eta,$$

where the threshold τ_η can be calibrated from held-out native samples. In our experiments, we report ROC-based AUC and TPR at target FPR levels using the corresponding Native-score distribution. FFR provides a matched linear readout for the SAC-induced evidence geometry under the stated calibration model. Appendix B.1–B.4 provide the full FFR rationale, matched-direction derivation, and relation to the old detector.

Furthermore, as one step towards robust detection, we introduce another detector, GO-FFR, extended from FFR by considering controlled insertion/deletion attacks:

$$Z_{\text{GO-FFR}}(Y) = \max_{d \in \mathcal{D}} Z_d(Y),$$

where $\mathcal{D}$ is a scan window and $Z_d(Y)$ is a test statistic defined in Appendix B.5. It scans a small family of global offsets to partially recover mismatched token–signal alignment, reuses the FFR scoring direction, and calibrates its threshold on native samples under the same offset-scan rule. Appendices B.5–B.6 give the full analyses.

3 Why Correlated Perturbations Help DLMs

This section provides a mechanism-level rationale for why smooth, locally correlated perturbations may be better aligned with DLM refinement dynamics than high-frequency i.i.d. perturbations. We analyze the geometry of the latent perturbation field and the induced detector evidence, and derive empirical predictions on generation quality and detectability that are evaluated in Section 4.

3.1 DLM Decoding Couples Neighboring Decisions

Autoregressive watermarking perturbs a left-to-right sampler whose prefix is fixed once a token is emitted. DLM decoding is different. LLaDA-style generation repeatedly predicts multiple masked positions in parallel, accepts or remasks tokens according to confidence, and revises a partially generated sequence over denoising steps (Nie et al., 2025; Ghazvininejad et al., 2019). More recent masked-diffusion and non-Markovian discrete-diffusion work likewise emphasizes iterative denoising and refinement-sensitive generation dynamics (Zhang et al., 2025; Sahoo et al., 2024). As a result, neighboring token positions are not isolated one-step decisions: local choices can be compared, retained, or revised together during refinement.

This difference makes the geometry of the perturbation field relevant. Beyond its marginal strength at each position, a watermark also determines how perturbations vary across adjacent positions. An i.i.d. Gumbel field can vary sharply across neighboring masked positions that are refined together, creating a plausible mismatch with the local consistency preference of iterative masked refinement. A smoother locally correlated field makes nearby perturbation preferences more coherent, which motivates the quality-stability hypothesis tested in Section 4. Figure 1 illustrates this contrast, while the following subsections formalize it using a latent local-difference proxy.

3.2 Independent Latent Fields Create Local Roughness

We first define a latent local-difference proxy before the coordinate-wise Gumbel transform. Fix a vocabulary item j, and consider the Gaussian latent field $Z(t, j)$ that is later mapped monotonically into Gumbel noise. We measure lag-one local variation by

$$D = \mathbb{E}\Big[(Z(t+1, j) - Z(t, j))^2\Big],$$

A larger D indicates sharper adjacent changes in the latent perturbation field.

For an i.i.d. latent Gaussian field, $\mathrm{Var}(Z(t, j)) = 1$ and $\mathrm{Cov}(Z(t, j), Z(t+1, j)) = 0$. Therefore, the difference proxy for i.i.d. latent Gaussian field is

$$D_{\text{i.i.d}} = 2.$$

3.3 SAC-Copula Reduces Local Perturbation Variation

SAC-Copula changes the joint structure of the perturbation field rather than the one-coordinate

marginal. Let $r_\tau(W, \sigma_K)$ denote the finite-window autocorrelation of the normalized spatial branch defined in Appendix A.3. This quantity is large at small lags when the smoothing kernel strongly overlaps with its shifted copy, and decays as the lag exceeds the kernel support. For the final latent field, the nonzero-lag covariance is

$$\mathrm{Cov}(Z_{\mathrm{final}}(t,j), Z_{\mathrm{final}}(t+\tau,j)) = \rho^2 r_\tau(W,\sigma_K).$$

Let $r_1 = r_{\tau=1}(W,\sigma_K)$, and write $\Delta_t = Z_{\mathrm{final}}(t+1,j) - Z_{\mathrm{final}}(t,j)$, the lag-one local difference energy becomes

$$D_{\mathrm{SAC}} = \mathbb{E}[\Delta_t^2] = 2(1-\rho^2 r_1).$$

When $r_1 > 0$ and $\rho > 0$, SAC-Copula has lower latent local variation than the i.i.d. field:

$$D_{\mathrm{SAC}} < D_{\mathrm{i.i.d}}.$$

This lower local variation gives a plausible quality mechanism for DLM decoding. LLaDA-style models refine neighboring masked positions jointly, so when clean logits are locally uncertain, abrupt independent perturbation changes can push nearby positions toward inconsistent random preferences. SAC-Copula does not reduce the single-coordinate probability of large Gumbel values: after the monotone Gaussian-copula-to-Gumbel transform, each coordinate still has the standard Gumbel marginal under the assumptions in Appendix A.2. Instead, it reduces abrupt local changes across positions by inducing positive low-lag dependence. This motivates the empirical hypothesis that smoother perturbation fields may improve quality stability under iterative masked refinement.

Because the inverse-Gumbel transform is nonlinear, this latent local-difference proxy characterizes only the dependence structure before the copula transform and should not be interpreted as invariance of the full generated-text distribution.

Figure 3 visualizes these perturbation-field diagnostics. The autocorrelation view shows that SAC-Copula introduces a positive low-lag dependence, while the adjacent-variation view shows the corresponding reduction in the latent roughness proxy.

Together, these results suggest that locally correlated perturbations better support DLM decoding quality while preserving watermark detectability. Section 4 evaluates these predictions empirically.

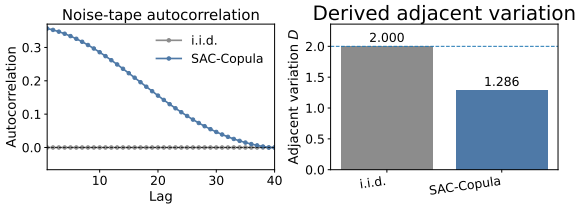


Figure 3: Perturbation-field smoothness diagnostics. SAC-Copula induces positive low-lag autocorrelation and lowers the latent adjacent-variation proxy from $D_{\mathrm{iid}} = 2$ to $D_{\mathrm{SAC}} \approx 1.286$. These are latent-field diagnostics, not text-quality measurements.

4 Experiments

We evaluate four questions: clean quality–detectability and matched-i.i.d. tail stability; correlation strength; detector, calibration, and targeted-transfer controls; and residual evidence under controlled token edits.

4.1 Experimental Setup

We evaluate generation from LLaDA-8B-Instruct (Nie et al., 2025) on ELI5 prompts (Fan et al., 2019). LLaDA–ELI5 is our primary benchmark; the targeted Dream-7B and C4-en evaluations reported below test transfer across a second DLM backbone and a second task/source setting without redefining the primary protocol. The comparison includes native decoding and the direct i.i.d. Gumbel baseline. We evaluate the LLaDA-adapted KGW-style (Kirchenbauer et al., 2023) and Unigram (Zhao et al., 2024) baselines, together with a PatternMark baseline (Chen et al., 2025), over $\delta \in \{1,2,3\}$. For KGW and Unigram, we use green-list fraction $\gamma = 0.25$; each method uses its mechanism-specific detector.

Unless otherwise stated, SAC-Copula uses $W = 39, \sigma_K = 15.0, \rho = 0.6$, controlling correlation range, kernel shape, and correlated-branch strength. The original mixed-parameter study informed this submitted setting (Appendix C.5); the fixed-(W,σ_K) sweep below directly tests ρ and supports $\rho = 0.6$ as a balanced interior operating point. Full generation, watermark, detector, calibration, seed, and evaluator details are reported in Appendix C.2.

Quality metrics (details in Table 6) include median PPL, PPL$>$ 100, SBERT, MAUVE (Pillutla et al., 2021), Token TV, Rep3, and Distinct3; the matched i.i.d.–SAC comparison additionally reports PPL tail collapse. AUC and TPR at nominal 1%/5% FPR summarize score discrimination (Li

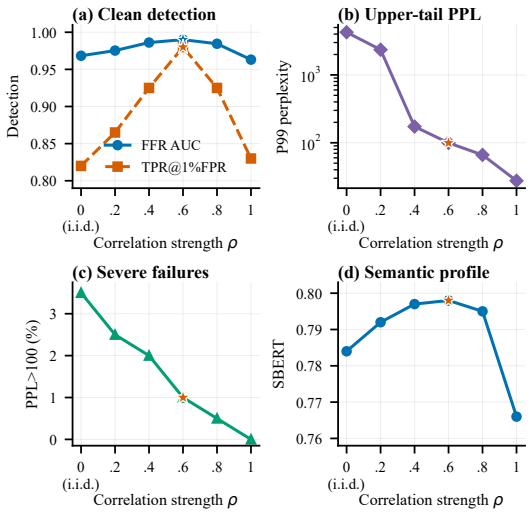


Figure 4: Correlation-strength ρ sweep under ($W = 39, \sigma_K = 15, \alpha = 1$); $\rho = 0$ is matched i.i.d. Gumbel. Detection peaks at moderate correlation and falls at $\rho = 1$; the PPL tail keeps improving, but broader endpoint quality degrades. $\rho = 0.6$ is a balanced interior point.

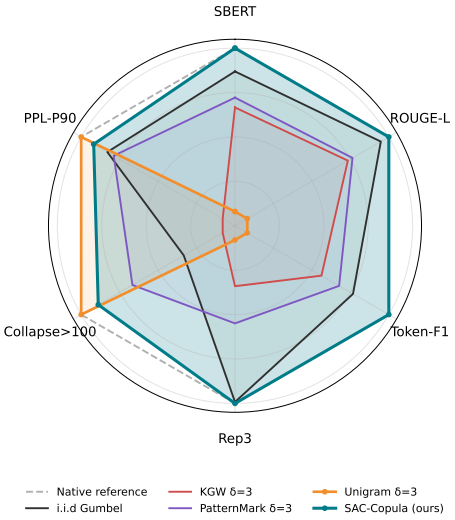


Figure 5: Generation-quality profile for SAC-Copula, i.i.d. Gumbel, and representative external baselines on direction-normalized axes; larger values indicate better or more native-like behavior.

et al., 2025; He et al., 2025); realized FPR under frozen thresholds tests operational transfer.

Ablation on Correlation Strength. Figure 4 varies only ρ under fixed $W = 39, \sigma_K = 15, \alpha = 1$, with $\rho = 0$ representing the i.i.d. Gumbel. FFR AUC/TPR@1%FPR rises from $0.9683/0.820$ at $\rho = 0$ to $0.9900/0.980$ at 0.6, then falls to $0.9631/0.830$ at 1. P99 decreases from 4256.63 to 99.65 and 27.55, but $\rho = 1$ worsens semantic, repetition, and diversity metrics relative to 0.6. Thus $\rho = 0.6$ is a balanced interior point; full tables are in Appendix C.9.

4.2 Text Quality and Stability

Text quality is the primary empirical target. Relative to matched i.i.d. Gumbel, the central PPL values are close (median 8.446 vs. 8.260), whereas the upper tail differs substantially: P95 decreases from 48.420 to 30.810, P99 from 4256.634 to 99.653, and the fraction above PPL 100 from 3.5% (7/200) to 1.0% (2/200). Thus, the clearest PPL gain is reduced severe upper-tail failures rather than a uniform per-prompt improvement. Table 1 and Figure 5 provide the broader context: SAC retains strong low-FPR detectability together with a competitive semantic, distributional, repetition, diversity, and token-drift profile. Full tail diagnostics and complete sweep results are in Appendix C.9.

4.3 Clean Detection and Calibration

Table 1 establishes clean fixed-alignment detectability; Appendix C.1, Figure 8, retains the full external sweep.

Detector-readout control. Under the legacy readout, i.i.d. versus SAC AUC/TPR@5%FPR is $0.966/0.855$ versus $0.987/0.950$; FFR further raises SAC TPR@1%FPR/TPR@5%FPR to $0.980/0.985$. Thus the SAC signal difference is visible under the common legacy readout, while FFR provides an additional covariance-aware readout gain. Appendix C.6, Table 15, gives the full control.

Calibration sensitivity. Table 2 condenses the model-native H_0 size controls from Appendix C.10: score discrimination remains strong and improves with N, while realized FPR at frozen nominal thresholds is operationally sensitive to calibration size. The four-source control in Table 3 shows only modest changes in score discrimination under this fixed protocol, quantifying calibration-source sensitivity. Threshold portability is nevertheless source-sensitive: Wikipedia-H_0 thresholds yield realized FPRs of 6.5% and 21.0% on ELI5 Native at nominal 1% and 5%, respectively, despite strong score discrimination. Freezing the original LLaDA–ELI5 thresholds without C4 refitting yields 0.95%/5.50% realized FPR on 2,000 held-out C4 `realnewslike` negatives at nominal 1%/5%; this is detector-only threshold transfer,

Method	Strength	AUC ↑	TPR@1%FPR ↑	TPR@5%FPR ↑	Median PPL ↓	PPL> 100 ↓	SBERT ↑	MAUVE ↑	Token TV ↓	Rep3 ↓	Distinct3 ↑	Quality ↑
Native	–	–	–	–	6.381	0.005	1.000	1.000	0.000	0.406	0.594	0.960
i.i.d. Gumbel	–	0.966	0.815	0.855	8.446	0.035	0.784	0.991	0.163	0.415	0.585	0.661
KGW	$\delta = 1$	0.670	0.060	0.120	6.697	0.030	0.794	**0.999**	**0.146**	0.414	0.586	0.694
KGW	$\delta = 2$	0.820	0.200	0.315	8.020	0.045	0.776	0.998	0.173	0.450	0.550	0.553
KGW	$\delta = 3$	0.936	0.515	0.660	10.543	0.050	0.762	0.982	0.228	0.510	0.489	0.379
Unigram	$\delta = 1$	0.896	0.385	0.555	**6.145**	**0.000**	**0.803**	0.981	0.233	**0.395**	**0.605**	**0.798**
Unigram	$\delta = 2$	0.988	0.850	0.930	6.535	**0.000**	0.755	0.841	0.360	0.455	0.545	0.545
Unigram	$\delta = 3$	**0.997**	**0.985**	**0.995**	6.585	0.005	0.690	0.601	0.439	0.560	0.440	0.150
PatternMark	$\delta = 1$	0.818	0.510	0.625	6.766	0.010	0.781	0.983	0.153	0.426	0.574	0.721
PatternMark	$\delta = 2$	0.926	0.760	0.800	7.263	0.015	0.779	0.996	0.161	0.422	0.578	0.714
PatternMark	$\delta = 3$	0.938	0.760	0.805	8.938	0.020	0.768	0.982	0.198	0.480	0.520	0.555
SAC-Copula (Ours)	–	0.990	0.980	0.985	8.260	0.010	0.798	**0.999**	0.163	0.414	0.586	0.758

Table 1: Clean detection and text-quality trade-off. AUC and low-FPR TPR report detection; Median PPL and PPL> 100 report central fluency and severe tail failures; the remaining raw metrics report broader quality. The normalized Quality score is descriptive only. **Best** and second-best watermarked results are marked in each column.

N	AUC	TPR@1%FPR	FPR@nom. 1%	FPR@nom. 5%
50	.9758±.0111	.8025±.0715	28.88%	34.33%
100	.9867±.0034	.9078±.0473	13.03%	19.55%
200	.9888±.0009	.9487±.0277	5.13%	11.15%
500	.9900	.9800	0.50%	8.50%

Table 2: Calibration-size N controls. AUC and TPR@1% are mean±SD for repeated $N < 500$ subsamples; the last columns report realized FPR at nominal 1%/5% thresholds.

H_0 source	AUC	TPR@1%FPR	TPR@5%FPR
Native	0.9900	0.980	0.985
ELI5 Human	0.9878	0.955	0.980
C4	0.9871	0.955	0.975
Wikipedia	0.9861	0.950	0.970

Table 3: Four-source H_0 calibration control at $N = 500$ with fixed development and evaluation identities.

not C4 generation. Full size, source, prompt, and threshold controls are in Tables 23–28.

Targeted model/task transfer. Table 4 evaluates SAC-Copula on two additional model–dataset settings. On Dream-7B/ELI5, SAC-Copula maintains near-perfect detection while substantially reducing P99 and the Composite-collapse rate. On LLaDA/C4-en, it improves detection while reducing P99 and the PPL-tail failure rate, with nearly unchanged median PPL (13.71 vs. 13.73). These targeted checks extend the evidence to a second backbone and task/source setting without establishing universal transfer. Appendix C.11 provides full Dream-7B/ELI5 (Table 29) and LLaDA/C4-en results, including a separate common-detector control (Tables 30–31).

Model/ Dataset	Method	AUC ↑	TPR@1%FPR ↑	P99 ↓	Tail failure ↓
Dream-7B/ ELI5	i.i.d. Gumbel	1.0000	1.000	5988.25	24.5%
	SAC-Copula	0.9986	0.995	240.68	4.0%
LLaDA/ C4-en	i.i.d. Gumbel	0.8260	0.415	35740.32	10.5%
	SAC-Copula	0.9138	0.670	494.57	3.0%

Table 4: Generalization across models and datasets. Tail failure denotes Composite collapse for Dream-7B/ELI5 and PPL >100 for LLaDA/C4-en

4.4 Residual Evidence under Controlled Token Edits

Figure 6 reports rate-dependent low-FPR evidence under controlled deletion, insertion, and substitution for i.i.d. Gumbel, PatternMark, SAC+FFR, and SAC+GO-FFR. Under deletion and insertion, GO-FFR retains more low-FPR evidence than fixed-alignment FFR once edits are applied, while PatternMark provides an external cross-method reference. Under substitution, the two SAC readouts coincide because positional alignment is preserved. Full AUC/TPR sweeps and additional selected operating-point diagnostics are reported in Appendix C.7.

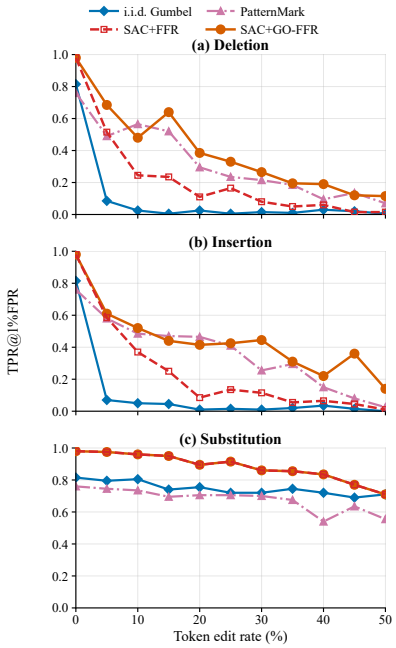


Figure 6: TPR at 1% FPR across controlled 0–50% token-edit sweeps for i.i.d. Gumbel, PatternMark, SAC+FFR, and SAC+GO-FFR. GO-FFR provides coarse synchronization recovery under deletion/insertion; under substitution, the two SAC readouts coincide.

GO-FFR scans a coarse global-offset family, while cumulative token–tape drift under deletion and insertion remains unresolved. Semantic/document paraphrasing replaces, reorders, or regenerates the evidence-bearing sequence and remains outside the current robustness scope. Full failure diagnostics are in Appendix C.8.

5 Discussion

SAC-Copula focuses on smooth, locally correlated DLM Gumbel fields and achieves a strong quality–detectability trade-off. Relative to matched i.i.d. Gumbel, its clearest quality gain is improved upper-tail stability with a competitive broader profile. The latent-field analysis is a mechanism-level rationale: one-coordinate marginals neither imply invariance of the complete generated-text distribution nor prove downstream text quality.

FFR provides a SAC-aware covariance-matched readout under the Gaussian mean-shift rationale. GO-FFR extends this readout with coarse global-offset recovery for insertion/deletion.

The size/source controls and frozen-threshold C4 test quantify calibration sensitivity and threshold transfer. LLaDA–ELI5 remains the primary setting, while Dream-7B–ELI5 and LLaDA–C4-en extend the evaluation to a second DLM backbone and a second task/source setting. The fixed-W, σ_K six-point sweep identifies an interior operating region around $\rho = 0.6$.

6 Limitations

Detection relies on calibrated H_0 samples and development data. Deployment requires private-tape access and representative calibration data. Natural extensions include broader architectural coverage, more systematic W/σ_K studies, runtime characterization, and deployment-oriented robustness evaluation (Zhao et al., 2025).

Under edit stress tests, GO-FFR partially recovers coarse insertion/deletion offsets, while low-FPR detection degrades as cumulative drift grows. Paraphrastic rewriting remains challenging because it breaks the fixed tape readout (Krishna et al., 2023; Hou et al., 2024). Broader watermark-security work studies strong-watermark impossibility limits and watermark stealing (Zhang et al., 2024; Jovanović et al., 2024).

References

Scott Aaronson. 2023. Watermarking of large language models. https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17. Accessed: 2023-08.

Avi Bagchi, Akhil Bhimaraju, Moulik Choraria, Daniel Alabi, and Lav R. Varshney. 2026. Watermarking discrete diffusion language models. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI - Interpretability, Robustness, and Safety across Modalities*.

Ruibo Chen, Yihan Wu, Yanshuo Chen, Chenxi Liu, Junfeng Guo, and Heng Huang. 2025. A watermark for order-agnostic language models. In *International Conference on Learning Representations*.

Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1125–1139. PMLR.

Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. 2023. Publicly-Detectable Watermarking for Language Models. Cryptology ePrint Archive, Paper 2023/1661.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. In *Advances in Neural Information Processing Systems 37*.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. 2026. Watermarking diffusion language models. In *International Conference on Learning Representations*.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2025. Theoretically grounded framework for llm watermarking: A distribution-adaptive approach. In *Advances in Neural Information Processing Systems 38*, pages 114528–114564.

Pyo Min Hong and Albert No. 2026. dgMARK: Decoding-guided watermarking for diffusion language models. In *Forty-third International Conference on Machine Learning*.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024. Semstamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4067–4082, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 22570–22593. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. 2025. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *Advances in Neural Information Processing Systems 38*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34*.

Ofek Raban, Gal Chechik, and Ethan Fetaya. 2026. Lrdwm: Efficient watermarking for diffusion language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43510–43517, San Diego, California, United States. Association for Computational Linguistics.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems 37*.

Chenrui Wang, Junyi Shu, Billy Chiu, Yu Li, Saleh Alharbi, Min Zhang, and Jing Li. 2025. Learning to watermark: A selective watermarking framework for large language models via multi-objective optimization. In *Advances in Neural Information Processing Systems 38*.

Linyu Wu, Linhao Zhong, Wenjie Qu, Yuexin Li, Yue Liu, Shengfang Zhai, Chunhua Shen, and Jiaheng Zhang. 2025. Dmark: Order-agnostic watermarking for diffusion large language models. *Preprint*, arXiv:2510.02902.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4031–4055, Mexico City, Mexico. Association for Computational Linguistics.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2024. Watermarks in the sand: Impossibility of strong watermarking for language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 58851–58880. PMLR.

Yangtian Zhang, Sizhuang He, Daniel Levine, Lawrence Zhao, David Zhang, Syed A Rizvi, Shiyang Zhang, Emanuele Zappala, Rex Ying, and David van Dijk. 2025. Non-markovian discrete diffusion with causal language models. In *Advances in Neural Information Processing Systems 38*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2024. Provable robust watermarking for ai-generated text. In *International Conference on Learning Representations*.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. 2025. Sok: Watermarking for ai-generated content. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2621–2639. IEEE.

A Watermark Properties

This appendix gives proof details behind the SAC-Copula field in Section 2.1. The claims are about the latent watermark tape. They are not claims about invariance of the complete output-text law.

A.1 Boundary Convention

The current convolution convention uses zero padding and divides by the global full-window energy

$$S_2 = \sum_{k=-h}^{h} K_W[k]^2.$$

Therefore exact unit variance is guaranteed for full-window interior positions. Boundary positions can have smaller variance because fewer nonzero taps contribute.

An all-position idealization would define

$$\mathcal{K}_t = \{k \in [-h, h] : 1 \le t - k \le L_{\max}\},$$
$$S_2(t) = \sum_{k \in \mathcal{K}_t} K_W[k]^2,$$

and divide the boundary convolution by $\sqrt{S_2(t)}$. In this paper, strict marginal claims are made under the full-window or boundary-normalized convention. Boundary deviations under zero padding are treated as implementation checks rather than as part of the exact marginal claim.

A.2 Marginal Gaussian and Gumbel Exactness

At a full-window position,

$$Z_{\text{spatial}}(t, j) = \sum_{k=-h}^{h} K_W[k] Z_{\text{white},1}(t - k, j).$$

This is a linear combination of independent Gaussian variables, hence Gaussian. Since the source variables have unit variance,

$$\text{Var}(Z_{\text{spatial}}(t, j)) = \sum_{k=-h}^{h} K_W[k]^2 = S_2.$$

Thus

$$\widetilde{Z}_{\text{spatial}}(t, j) = \frac{Z_{\text{spatial}}(t, j)}{\sqrt{S_2}} \sim \mathcal{N}(0, 1).$$

The same proof applies to the boundary-normalized ideal form after replacing S_2 by $S_2(t)$.

The white stream and spatial stream are independent. Therefore

$$Z_{\text{final}}(t, j) = \sqrt{1 - \rho^2} Z_{\text{white},2}(t, j) + \rho \widetilde{Z}_{\text{spatial}}(t, j)$$

is Gaussian with zero mean and variance $(1 - \rho^2) + \rho^2 = 1$. Hence each valid coordinate of Z_{final} is $\mathcal{N}(0, 1)$.

By the probability integral transform,

$$U(t, j) = \Phi(Z_{\text{final}}(t, j)) \sim \text{Uniform}(0, 1).$$

Applying the inverse standard Gumbel CDF gives

$$G_{\text{corr}}(t, j) = F_G^{-1}(U(t, j)) \sim \text{Gumbel}(0, 1).$$

For the standard Gumbel distribution, $F_G^{-1}(u) = -\log(-\log u)$. In numerical evaluation, U is clamped to avoid floating-point extremes, so exact unbounded Gumbel statements refer to the ideal transform.

A.3 Finite-Window Latent Autocorrelation

Let

$$b_k = \frac{K_W[k]}{\sqrt{S_2}}, \qquad b_k = 0 \quad \text{for } k \notin [-h, h].$$

Then

$$\widetilde{Z}_{\text{spatial}}(t, j) = \sum_k b_k Z_{\text{white},1}(t - k, j).$$

Expanding the covariance and using independence of the white source gives the finite-window autocorrelation

$$r_\tau(W, \sigma_K) = \sum_k b_k b_{k+\tau}.$$

Thus $r_0 = 1$, $r_{-\tau} = r_\tau$, and $r_\tau = 0$ for $|\tau| > W - 1$. For the final latent field and $\tau \neq 0$,

$$\text{Cov}(Z_{\text{final}}(t, j), Z_{\text{final}}(t + \tau, j)) = \rho^2 r_\tau(W, \sigma_K),$$

because the white stream contributes no cross-position covariance. This latent autocorrelation motivates the detector-side H_ω template, while the detector covariance itself is estimated from extracted Native evidence.

For the lag-one local-difference proxy used in Section 3.3, the unit marginal variance and the covariance above give

$$D_{\text{SAC}} = \mathbb{E}\Big[(Z_{\text{final}}(t + 1, j) - Z_{\text{final}}(t, j))^2\Big] = 2 - 2\rho^2 r_1(W, \sigma_K) = 2(1 - \rho^2 r_1).$$

Thus the Section 3.3 expression is a direct consequence of the finite-window autocorrelation and the preserved one-coordinate latent variance.

A.4 Latent Tape KL Proxy

Let $P_0 = \mathcal{N}(0, I_T)$ be an i.i.d. latent Gaussian tape and $P_\rho = \mathcal{N}(0, \Sigma_T)$ the SAC latent Gaussian tape, where

$$\Sigma_T = (1-\rho^2)I_T + \rho^2 R_T,$$
$$[R_T]_{ab} = r_{a-b}(W, \sigma_K).$$

Assume $\Sigma_T \succ 0$, so the log-determinant is well-defined. This holds for $\rho < 1$ when R_T is positive semidefinite; the endpoint $\rho = 1$ additionally requires nonsingularity of R_T.

For zero-mean Gaussians,

$$D_{\mathrm{KL}}(P_\rho \parallel P_0) = \frac{1}{2}\left(\operatorname{tr}\Sigma_T - T - \log\det\Sigma_T\right).$$

Since $\operatorname{diag}(\Sigma_T) = 1$, $\operatorname{tr}\Sigma_T = T$, and

$$\mathcal{L}^{(T)}_{\mathrm{tape}} = \frac{1}{T} D_{\mathrm{KL}}(P_\rho \parallel P_0) = -\frac{1}{2T}\log\det\Sigma_T.$$

This is a latent process diagnostic, not semantic KL and not output-text KL.

This expansion is an explanatory approximation obtained from the log-determinant series around the identity covariance. Set $A = R_T - I_T$. Since $\Sigma_T = I_T + \rho^2 A$ and $\operatorname{tr}(A) = 0$,

$$\mathcal{L}^{(T)}_{\mathrm{tape}} = \frac{\rho^4}{4T}\operatorname{tr}(A^2) + O(\rho^6).$$

For the Toeplitz finite-window case,

$$\operatorname{tr}(A^2) = \sum_{\tau \neq 0}(T - |\tau|) r_\tau(W, \sigma_K)^2.$$

The leading term is $O(\rho^4)$ because the diagonal marginal variances are preserved and the trace term cancels.

A.5 Implementation Checks

Implementation checks should include:

- marginal Gaussian and Gumbel QQ plots;
- empirical versus theoretical autocorrelation;
- low-frequency energy of the noise tape;
- effective sample size estimates;
- boundary variance curves.

For a diagnostic sequence a, with discrete Fourier transform $\widehat{a}(f)$, define the low-frequency energy fraction as

$$\mathrm{LFE}(a) = \frac{\sum_{f \in \mathcal{F}_{\mathrm{low}}} |\widehat{a}(f)|^2}{\sum_{f \in \mathcal{F}} |\widehat{a}(f)|^2}.$$

These checks validate the noise-tape construction, not the distributional invariance of generated text.

B Detector Rationale

This appendix expands the detector rationale in Section 2.3. It clarifies why the detector is matched to the evidence geometry induced by the SAC-Copula perturbation field.

B.1 From i.i.d. Evidence to Covariance-Aware Calibration

The old detector aggregates raw evidence with an equal-weight i.i.d. Z-score:

$$Z_{\mathrm{old}} = \frac{\mathbf{1}^\top(\mathbf{s} - \mu_G \mathbf{1})}{\sigma_G\sqrt{T}}.$$

It implicitly assumes

$$\Sigma_{x,0} \approx \sigma_G^2 I, \qquad \Delta_x \propto \mathbf{1}.$$

For non-i.i.d. evidence, however, the null variance of the equal-weight sum is

$$\mathrm{Var}_0(\mathbf{1}^\top \mathbf{x}) = \mathbf{1}^\top \Sigma_{x,0} \mathbf{1}.$$

Positive local dependence increases this variance relative to the i.i.d. value, reducing effective sample size. HAC or covariance-aware calibration corrects this null variance while keeping the evidence direction $\mathbf{1}$. If a calibration-only score is a positive rescaling,

$$Z_{\mathrm{h0_hac}} = c_d Z_{\mathrm{old}}, \qquad c_d > 0,$$

then within the same detector group and fixed calibration setting it does not change ROC ordering or AUC. This explains why calibration alone is insufficient when the useful watermark shift is not aligned with $\mathbf{1}$.

B.2 Why Filtering is Needed for SAC-Copula Evidence

SAC-Copula injects a locally correlated, low-lag perturbation field. The resulting evidence is not merely a uniform mean shift across positions. Full Filtered Ridge maps centered raw evidence into a SAC-aligned filtered feature:

$$\mathbf{u} = H_\omega \mathbf{x}.$$

The filter H_ω is motivated by the autocorrelation envelope of the SAC kernel, using weights proportional to $\exp(-\tau^2/(4\sigma_K^2))$ with boundary renormalization. Concretely, with

$$a_\tau = \exp\left(-\frac{\tau^2}{4\sigma_K^2}\right)\mathbf{1}\{|\tau| \le B\},$$

a row-normalized filtered readout can be written as

$$[H_\omega]_{t,s} = \frac{a_{s-t}}{\sum_{r:\,1\le t+r\le T} a_r}.$$

where $a_r = 0$ outside $[-B, B]$. This filter is a detector-side transform, not generation noise.

Native text and the decoding process may also introduce background correlations. Therefore the latent tape covariance should not be used directly as the detector's null covariance. The detector estimates the observed-evidence Native covariance and projects it into filtered space:

$$\widehat{\Sigma}_{u,0} = H_\omega \widehat{\Sigma}_{x,0} H_\omega^\top.$$

Thus FFR combines a SAC-motivated feature map with Native H_0 observed-evidence covariance whitening.

B.3 Ridge-Regularized Matched Direction

In the detector statistic, raw evidence is centered before filtering, so $\mathbf{u} = H_\omega \mathbf{x}$ is already H_0-centered up to finite-sample error. The surrogate $T_w(\mathbf{u}) = w^\top(\mathbf{u} - \boldsymbol{\mu}_{u,0})$ writes the same idea in abstract filtered-space notation.

In filtered feature space, use the Gaussian mean-shift surrogate

$$\begin{aligned} H_0 :&\quad \mathbf{u} \sim (\boldsymbol{\mu}_{u,0}, \Sigma_{u,0}),\\ H_1 :&\quad \mathbf{u} \sim (\boldsymbol{\mu}_{u,0} + \boldsymbol{\Delta}_u, \Sigma_{u,0}). \end{aligned}$$

For a linear score

$$T_w(\mathbf{u}) = w^\top(\mathbf{u} - \boldsymbol{\mu}_{u,0}),$$

the standardized separation is

$$\mathrm{SNR}(w) = \frac{w^\top \boldsymbol{\Delta}_u}{\sqrt{w^\top \Sigma_{u,0} w}}.$$

Let $b = \Sigma_{u,0}^{1/2} w$, so $w = \Sigma_{u,0}^{-1/2} b$, and

$$a = \Sigma_{u,0}^{-1/2} \boldsymbol{\Delta}_u.$$

Then

$$\begin{aligned} \mathrm{SNR}(w) &= \frac{b^\top a}{\|b\|_2}\\ &\le \|a\|_2. \end{aligned}$$

Equality holds when $b \propto a$, giving

$$w^\star \propto \Sigma_{u,0}^{-1} \boldsymbol{\Delta}_u.$$

The practical direction is the ridge-regularized plug-in estimate

$$w_\lambda = (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \widehat{\boldsymbol{\Delta}}_u.$$

Because the main detector centers evidence in raw space before filtering, the implemented scalar score can be written without an additional filtered-space mean subtraction:

$$Z_{\mathrm{FFR}}(y) = \frac{\widehat{\boldsymbol{\Delta}}_u^\top (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \mathbf{u}(y)}{\gamma_{\mathrm{FFR}}}.$$

For scores expressed directly in centered raw-evidence space, the corresponding linear direction is

$$\begin{aligned} \mathbf{q} &= \frac{H_\omega^\top w_\lambda}{\gamma_{\mathrm{FFR}}},\\ \gamma_{\mathrm{FFR}} &= \sqrt{\widehat{\boldsymbol{\Delta}}_u^\top (\widehat{\Sigma}_{u,0} + \lambda I)^{-1} \widehat{\boldsymbol{\Delta}}_u}. \end{aligned}$$

The ridge is finite-sample regularization for an empirical covariance inverse. The direction $\widehat{\boldsymbol{\Delta}}_u$ must be estimated from a development split or cross-fitting protocol without evaluation leakage. The optimality statement is limited to the linear detector class under the shared-covariance Gaussian surrogate.

B.4 Relation to the Original Detector

FFR reduces to the old detector under

$$\begin{aligned} H_\omega &= I, & \widehat{\Sigma}_{u,0} &= \sigma_G^2 I,\\ \widehat{\boldsymbol{\Delta}}_u &\propto \mathbf{1}, & \lambda &= 0,\\ \widehat{\boldsymbol{\mu}}_{x,0} &= 0. \end{aligned}$$

In this case the FFR direction is proportional to $\mathbf{1}$, and the score becomes the old equal-weight Gumbel evidence sum. FFR is therefore a covariance-aware, direction-aware, SAC-aware generalization of the original i.i.d. detector.

B.5 Global Offset Extension

FFR assumes fixed token-to-tape alignment $j = i$. Insertion and deletion can instead induce

$$j = i + \delta(i),$$

where $\delta(i)$ is an unknown drift. For an alignment A, the ideal alignment-aware centered evidence would be

$$x_i^A = s(y_{A(i)}; i) - \mu_G - \widehat{\mu}_{x,0,i},$$

with score

$$Z(A) = \mathbf{q}^\top \mathbf{x}^A.$$

GO-FFR restricts alignment to the global-offset family

$$\mathcal{A}_{\mathrm{GO}} = \{A_d(i) = i + d : \ d \in \mathcal{D}\},$$

where $\mathcal{D} \subset \mathbb{Z}$ is a finite candidate offset set. In the reported attack-aware implementation, $\mathcal{D} = \{-D_{\max}, \ldots, 0\}$ for deletion, $\mathcal{D} = \{0, \ldots, D_{\max}\}$ for insertion, and $\mathcal{D} = \{0\}$ for fixed-alignment clean or substitution settings, with $D_{\max} = 96$. The resulting score is

$$Z_{\text{GO-FFR}}(Y) = \max_{d \in \mathcal{D}} Z_d(Y),$$
$$Z_d(Y) = \mathbf{q}^\top \mathbf{x}^{(d)}(Y).$$

This is a GLRT-style scan over a small alignment family. The max over offsets inflates the Native H_0 right tail. The GO-specific threshold is the empirical $1 - \eta$ quantile of $Z_{\text{GO-FFR}}$ on Native H_0 samples under the same offset-scan rule. GO-FFR recovers coarse global synchronization within the scanned global-offset family; cumulative local drift and semantic rewriting remain outside this alignment family.

B.6 Scope and Limitations

FFR adapts detection to correlated SAC-Copula evidence, while GO-FFR extends the same scoring direction with coarse synchronization recovery for mild insertion/deletion. Low-FPR evaluation uses detector-specific Native H_0 calibration, and $\widehat{\boldsymbol{\Delta}}_u$ is estimated without evaluation leakage. Semantic rewriting and cumulative edit-path drift remain outside the global-offset family analyzed here.

C Additional Experiment Evidence

This appendix records supporting clean-setting summaries, frozen-output re-analysis, calibration and targeted-transfer controls, and frozen attack-summary evidence used in Section 4. The clean quality and clean detection claims in the main text are supported by the frozen clean multi-baseline sweep summaries, not by the attack-summary evidence. The attack sources are summary-only artifacts derived from token-level edit outputs and semantic/paraphrase attack summaries; the attack-detail and failure-analysis sections do not use per-sample generated text, exact noise tapes, or full detector-score files.

C.1 Clean Quality and Detection Details

This subsection records supporting material for the clean multi-baseline comparison in Section 4.2 and Section 4.3. The source tables are frozen clean-text sweep summaries used to construct the main trade-off table, quality-score components, radar normalization, and heatmap points. These tables explain the compact quality score and visualization normalization; they do not introduce additional experimental claims beyond the clean-setting trade-off reported in the main text. Table 6 defines metric directions, Tables 9–10 report the full raw quality metrics, Tables 11–12 document the radar normalization, and Figures 7–8 provide compact clean quality–detection and clean detection views.

C.2 Full Experimental Configuration

This appendix records implementation-level details for reproducibility. Fields not exposed by the frozen scripts or summary metadata are marked as unavailable rather than inferred.

Baseline adaptation notes. KGW is adapted as a context-greenlist bias baseline with the configured context width, PRF, and detector listed in Table 7 and Table 8. Unigram is used as a static-greenlist clean-text trade-off baseline with its unique-token detector. PatternMark is the MarkovPattern ReferenceLite variant used in the clean evaluation, not a claim of a complete independent reimplementation of every PatternMark setting. Detector derivations for the old readout, covariance-aware calibration, filtering, ridge matched direction, and GO-FFR are provided in Appendix B.1, Appendix B.2, Appendix B.3, and Appendix B.5; this experiment appendix only records how those detectors are used in the experiments.

Quality–detection heatmap. Figure 7 summarizes the same clean-setting trends using a compact heatmap. The quality score averages direction-normalized components for semantic preservation, MAUVE-based distributional fidelity, token-distribution fidelity, low repetition, lexical diversity, and PPL-tail stability. This visualization is intended as a compact summary of the quality–detection balance rather than a replacement for the raw metric values in Table 9.

The external-baseline clean detection sweep over $\delta \in \{1, 2, 3\}$ is fully visualized below in Figure 8 and supports the clean multi-method comparison in Table 1. Figure 7 and Table 15 further connect

Evidence item	Appendix location	Source / scope note
Full experimental setup	Tables 7–8	implementation scripts, run metadata, and frozen summary files.
Metric definitions and orientation	Table 6	Defines raw metrics, direction, and visualization-only quality score.
Full clean quality metrics	Tables 9–10	frozen clean multi-baseline quality summaries.
Radar normalization	Tables 11–12	Display normalization only; not primary evidence.
Clean quality–detection view	Figure 7	Compact appendix view of clean quality and detection trade-offs.
Full clean detection sweep	Figure 8	External-baseline strength sweep supporting the clean multi-method comparison.
Unigram strength trade-off	Figure 9	Contrastive static-bias sweep.
SAC parameter sensitivity	Table 14	Raw-metric support for the selected SAC operating point.
Detector comparison	Table 15	Clean old-readout / FFR / GO-FFR sanity check.
Token-edit sweeps	Figures 12–14	AUC, TPR@FPR1, and TPR@FPR5 across 0–50% edit rates.
Failure analysis	Tables 17–18	Boundary evidence for edit drift and semantic rewriting.
Tail stability	Tables 19–20	Frozen-output re-analysis and paired upper-tail diagnostics.
Six-point ρ sweep	Tables 21–22	Fixed-parameter correlation-strength control.
Calibration controls	Tables 23–27	Size, source, threshold, and prompt-shift diagnostics.
Frozen C4 FPR	Table 28	Detector-only transfer of frozen ELI5 operating thresholds.
Dream-7B transfer	Table 29	Targeted second-backbone evidence on ELI5.
C4-en transfer	Table 30	Targeted method-matched continuation evidence.
C4 common-detector control	Table 31	Separate detector-control protocol for C4 generation.
Detector derivation	Appendix B	Full detector definitions are not duplicated in this experiment appendix.

Table 5: Experiment appendix evidence index. Each row points to the compact appendix evidence used to support Section 4. The index is organizational only and does not introduce new experimental values.

Metric	Definition / source	Direction	Use in this paper
PPL	Perplexity under the configured evaluator model.	Lower is better	Fluency and tail-stability evidence.
PPL-P90	90th-percentile conditional PPL.	Lower is better	Radar tail-stability axis.
Median / P95 / P99 PPL	Quantiles of the per-output conditional-PPL distribution.	Lower is better	Central and upper-tail summaries.
LLaDA PPL>100	Fraction of evaluated outputs with $\texttt{Conditional_PPL} > 100$.	Lower is better	LLaDA severe-tail diagnostic.
Dream Composite collapse	Preregistered Dream PPL/repetition criterion; distinct from LLaDA PPL>100.	Lower is better	Dream-specific collapse statistic.
SBERT	Sentence-embedding similarity to the Native reference.	Higher is better	Semantic preservation.
BERTScore	Token-level contextual similarity.	Higher is better	Semantic preservation in full quality table.
ROUGE-L	Longest-common-subsequence lexical overlap.	Higher is better	Radar lexical overlap axis.
Token-F1	Native token-set overlap F1.	Higher is better	Radar token-overlap axis.
MAUVE	Distributional similarity against Native reference generations.	Higher is better	Distributional fidelity.
Token TV / Token JS	Unigram token-distribution divergence from Native.	Lower is better	Token-distribution drift.
Top-k overlap	Overlap with Native top-k token set.	Higher is better	Token-distribution fidelity.
Rep3	Repeated 3-gram rate.	Lower is better	Repetition diagnostic.
Distinct3	Distinct 3-gram ratio.	Higher is better	Lexical diversity.
Ent3	3-gram entropy.	Higher is better	Diversity / spread diagnostic.
AUC	Threshold-independent detector ranking metric.	Higher is better	Clean and attack diagnostic detection.
TPR@FPR1 / TPR@FPR5	True-positive rate at 1% / 5% false-positive rate.	Higher is better	Low-FPR clean detection and attack diagnostics.
Quality	Direction-normalized visualization summary.	Higher is better	Descriptive radar/heatmap display only; not primary evidence.

Table 6: Metric definitions and orientations used in Section 4. Raw metrics are the primary evidence. The normalized Quality score is a descriptive display summary constructed from direction-normalized components and should not be interpreted as an additional ground-truth quality label.

the clean delta sweep to the raw quality tables and detector-readout sanity checks without duplicating the full CSV rows.

Full clean detection sweep. Figure 8 reports the full clean detection visualization for AUC, TPR@FPR1, and TPR@FPR5 across the external-baseline strength sweep. This appendix figure provides the full clean-detection sweep across methods and watermark-strength settings and does not introduce new experimental values.

C.3 Unigram Strength Trade-off

Figure 9 keeps the Unigram strength sweep as a contrastive reference for static-bias tuning. It is included in the appendix so that the main text can focus on the SAC-Copula clean quality and clean detection profile. The figure shows that stronger Unigram settings can improve clean detectability, but the same sweep changes semantic, distributional, token-drift, repetition, and diversity dimensions. This supports the main-text interpretation of Unigram as a trade-off baseline, not as the central object of study.

C.4 Traceable Clean-Generation Example

Table 13 provides one traceable qualitative example from the frozen generated-text JSONL artifact. The example is included only to make the automatic quality metrics more concrete; it is not used as primary evidence and is not selected to prove a general qualitative claim.

C.5 SAC Parameter Sensitivity

This subsection documents the SAC operating point used in the main experiments. The three SAC parameters have distinct roles: W controls the correlation window or range, σ_K controls the kernel decay and smoothness, and ρ controls the strength of the correlated branch. Table 14 compares several SAC configurations under the same clean-text evaluation slice.

Category	Field	Setting
Generation	Model name	`LLaDA-8B-Instruct`
Generation	Prompt source	`ELI5.jsonl`
Generation	Prompt format	LLaDA chat template
Generation	Split sizes	Native $H_0 = 500$, Delta-dev $= 200$, eval $= 200$
Generation	Operating GPU	NVIDIA RTX PRO 6000
Generation	Generation length	256 tokens
Generation	Block length	32
Generation	Diffusion steps	$\max(1, \lfloor 0.5\times$ generation length$\rfloor) = 128$
Generation	Temperature	1.0
Generation	top-p / top-k	not available in the frozen configuration fields
Generation	Mask id	126336
Generation	Seeds and key	fixed random seeds and a fixed private watermark key; exact values are kept in the implementation configuration
Generation	Cached artifacts	generated texts and exact perturbation tapes are cached for reproducibility in the implementation artifacts
Watermark	Main SAC-Copula	$W = 39, \sigma_K = 15.0, \rho = 0.6$
Watermark	SAC sensitivity rows	$W = 35, \sigma_K = 11.5, \rho = 0.7$; $W = 35, \sigma_K = 13.5, \rho = 0.7$; $W = 39, \sigma_K = 15.0, \rho = 0.6$
Watermark	Perturbation scale	$\alpha = 1.0$, passed as the generation temperature in the generation call
Watermark	Tape length	256 generated positions in the clean evaluation
Watermark	Exact tapes	saved under `exact_tapes`; missing-tape reconstruction disabled
Baselines	Direct mechanism baseline	i.i.d. position-indexed Gumbel
Baselines	External baselines	KGW context-green bias, Unigram static-green bias, PatternMark MarkovPattern ReferenceLite
Baselines	External sweep	$\delta \in \{1, 2, 3\}$ for KGW, Unigram, and PatternMark; $\gamma = 0.25$ for KGW and Unigram
Baselines	KGW details	conv-kernel PRF, kernel $[-1]$, context width 1, sumhash seeding, no-bias fallback
Baselines	Unigram details	static greenlist bias; unique-token detector is the main detector
Baselines	PatternMark details	MarkovPattern ReferenceLite, $k = 2$, pattern length 4, 10000 null samples

Table 7: Generation, watermark, and baseline configuration for the clean multi-baseline evaluation. Values are taken from the frozen generation/detection script and collected run metadata.

Category	Field	Setting
Detection	Canonical detector mapping	i.i.d. Gumbel: old z-score; KGW: KGWGreenlistZDetectorV2; Unigram: UnigramUniqueGreenlistDetectorV2; PatternMark: PatternMarkMarkovPatternDetector
Detection	SAC detector rows	old z-score, Full Filtered Ridge, Global Offset Full Filtered Ridge
Detection	FFR ridge	$\lambda = 10^{-4}$
Detection	Cross-fitting	2 folds with crossfit seed 20260506 for empirical direction estimation
Detection	Native H_0 calibration	detector-specific calibration on held-out Native H_0 samples
Detection	Target FPRs	1% and 5%
Detection	GO offset setting	max offset 96 and offset penalty 0.0; clean GO-FFR scoring uses offset 0
Detection	Tape device	generation tape on runtime device; detection tape on CPU
Quality	PPL model	Llama-3.1-8B-Instruct evaluator checkpoint
Quality	PPL collapse threshold	PPL > 100
Quality	PPL trim quantile	0.95
Quality	SBERT model	all-MiniLM-L6-v2 sentence-embedding model
Quality	BERTScore model	`roberta-large`, English, no baseline rescaling
Quality	MAUVE setup	Native reference group, `gpt2-large` features, max text length 256, 10 buckets
Quality	Token drift	unigram JS, TV, and top-k overlap for $k \in \{50, 100, 500, 1000\}$
Quality	Repetition/diversity	Rep3, Distinct3, Ent3, with Self-BLEU 3/4 available in multidimensional quality metadata
Quality	Normalized quality score	descriptive direction-normalized summary for visualization, not a primary metric

Table 8: Detector, calibration, and evaluator configuration for the clean multi-baseline evaluation. Values are taken from the implementation scripts, collected run metadata, and existing summary tables.

Detection metrics come from the FFR clean-detection summary, and quality metrics come from the multidimensional quality and token-drift summaries in the shared clean-text evaluation slice indexed as $\delta = 1$ for the external-baseline strength sweep. This δ index is not a SAC parameter; the SAC parameters are W, σ_K, ρ. The normalized quality score is not included in these delta-sweep summary files, so it remains marked as unavailable rather than reconstructed, and the interpretation should rely on the raw listed metrics.

The table should be read as a compact trade-off summary. The $W = 35, \rho = 0.7$ variants can improve some central fluency or collapse indicators, especially for $W = 35, \sigma_K = 13.5, \rho = 0.7$, but they also shift diversity, token-drift, and low-FPR detection behavior. The selected $W = 39, \sigma_K = 15.0, \rho = 0.6$ setting is used because it retains strong low-FPR detection while maintaining favorable MAUVE, repetition/diversity, and token-drift behavior.

Table 14 supports the selected configuration as a deliberate quality–detectability operating point over the reported local sensitivity configurations.

C.6 Detector Comparison for SAC Evidence

This subsection is included to separate generation-side structure from detector-side readout in the clean setting. The old detector is the i.i.d. equal-weight baseline readout, while FFR is the SAC-aware filtered statistic used to read the local low-lag evidence geometry induced by SAC-Copula. GO-FFR provides continuity with the mild insertion/deletion stress tests.

Table 15 reports the available clean detector comparison rows from the clean detection summary. The comparison should be read in two steps: if SAC evidence already improves over i.i.d. Gumbel under the old detector, that supports a generation-side signal difference; if FFR improves further on SAC evidence, that supports a matched readout for SAC structure.

Method	Strength	Mean PPL	Median PPL	Collapse	SBERT	BERTScore	MAUVE	Token TV
Native	–	16.840	6.381	0.005	1.000	1.000	1.000	0.000
i.i.d. Gumbel	–	636.069	8.446	0.035	0.784	0.868	0.991	0.163
KGW	$\delta = 1$	5.32×10^4	6.697	0.030	0.794	0.869	0.999	0.146
KGW	$\delta = 2$	6.98×10^4	8.020	0.045	0.776	0.867	0.998	0.173
KGW	$\delta = 3$	8.30×10^4	10.543	0.050	0.762	0.863	0.982	0.228
Unigram	$\delta = 1$	7.714	6.145	0.000	0.803	0.867	0.981	0.233
Unigram	$\delta = 2$	7.630	6.535	0.000	0.755	0.851	0.841	0.360
Unigram	$\delta = 3$	9.464	6.585	0.005	0.690	0.836	0.601	0.439
PatternMark	$\delta = 1$	3883.179	6.766	0.010	0.781	0.869	0.983	0.153
PatternMark	$\delta = 2$	3898.094	7.263	0.015	0.779	0.866	0.996	0.161
PatternMark	$\delta = 3$	3901.117	8.938	0.020	0.768	0.863	0.982	0.198
SAC-Copula	main	12.164	8.260	0.010	0.798	0.868	0.999	0.163

Table 9: Full automatic quality metrics for the clean multibase comparison, part 1. The table reports fluency and stability, semantic preservation, distributional fidelity, and token-distribution drift. Lower values are better for PPL, collapse, and token TV; higher values are better for semantic similarity and MAUVE.

Method	Strength	Token JS	Top50	Top100	Rep3	Distinct3	Ent3	Quality
Native	–	0.000	1.000	1.000	0.406	0.594	3.796	0.960
i.i.d. Gumbel	–	0.059	0.940	0.830	0.415	0.585	3.680	0.661
KGW	$\delta = 1$	0.057	0.920	0.900	0.414	0.586	3.798	0.694
KGW	$\delta = 2$	0.060	0.900	0.850	0.450	0.550	3.555	0.553
KGW	$\delta = 3$	0.068	0.860	0.830	0.510	0.489	3.179	0.379
Unigram	$\delta = 1$	0.074	0.860	0.760	0.395	0.605	3.967	0.798
Unigram	$\delta = 2$	0.120	0.660	0.600	0.455	0.545	3.819	0.545
Unigram	$\delta = 3$	0.164	0.600	0.430	0.560	0.440	3.408	0.150
PatternMark	$\delta = 1$	0.058	0.940	0.840	0.426	0.574	3.686	0.721
PatternMark	$\delta = 2$	0.061	0.900	0.800	0.422	0.578	3.809	0.714
PatternMark	$\delta = 3$	0.066	0.880	0.810	0.480	0.520	3.442	0.555
SAC-Copula	main	0.061	0.920	0.810	0.414	0.586	3.745	0.758

Table 10: Full automatic quality metrics for the clean multibase comparison, part 2. Lower values are better for token JS and repetition; higher values are better for vocabulary-overlap, distinctness, entropy, and the normalized quality score.

The clean fixed-alignment setting explains why GO-FFR matches or closely tracks FFR here: offset recovery is unnecessary when alignment is intact. Table 15 serves as a clean-detector sanity check connecting the shared legacy readout to the SAC-aware readout; GO-FFR is included for continuity with the controlled edit diagnostics.

C.7 Controlled Token-Level Edit Details

Table 16 and Figures 12–14 support Section 4.4. These results are intended as controlled token-edit diagnostics rather than a claim of adversarial robustness. Across selected deletion, insertion, and substitution settings, the SAC-aware detector retains partial residual evidence, and the Global Offset variant provides a coarse realignment benefit. The full sweep figures report AUC, TPR@FPR1, and TPR@FPR5 across 0–50% attack rates. Low-FPR TPR remains substantially harder than AUC under deletion and insertion, which supports the failure analysis rather than a broad edit-robustness claim. These conclusions come from the frozen token-edit summaries, relative-gain, and global-offset summaries. They should not be inferred from the multi-baseline clean-text results.

The full 20% cross-method AUC and detector/readout comparison is shown in Figure 11. Table 16 focuses on the detector-readout rows used to interpret SAC OLD, FFR, and GO-FFR behavior. Substitution does not induce cumulative position drift. Under substitution, the GO-FFR scan reduces to $\mathcal{D} = \{0\}$ and therefore coincides with fixed-alignment FFR; both curves are retained in the sweep figures for consistent cross-panel comparison.

Radar axis	Raw metric source	Original direction	Normalized score meaning
SBERT	SBERT	Higher is better	Embedding-level semantic closeness to Native text.
ROUGE-L	Mean ROUGE-L vs. Native	Higher is better	Lexical-sequence overlap with Native generations.
Token-F1	Native token-set F1	Higher is better	Native token-set overlap.
Rep3	Mean Rep3	Lower is better	Lower repeated 3-gram rate after inverted min–max normalization.
Collapse>100	CollapseRate_GT100	Lower is better	Fewer severe PPL-tail collapse cases after inverted min–max normalization.
PPL-P90	Conditional PPL P90	Lower is better	Lower 90th-percentile conditional perplexity after inverted min–max normalization.

Table 11: Metric orientation used in the v3.8 quality radar. All radar axes are direction-normalized so that larger values indicate better or more Native-like behavior. Lower-is-better metrics, including repetition, collapse rate, and PPL-P90, are inverted before normalization. The normalization is used only for visualization; raw quality metrics remain the primary evidence in Table 1, Table 9, and Table 10.

Method	SBERT	ROUGE-L	Token-F1	Rep3	Collapse>100	PPL-P90
Native	1.000	1.000	1.000	1.000	1.000	1.000
i.i.d. Gumbel	0.869	0.948	0.767	0.992	0.333	0.830
KGW $\delta = 3$	0.666	0.734	0.562	0.340	0.080	0.080
PatternMark $\delta = 3$	0.721	0.764	0.677	0.549	0.667	0.788
Unigram $\delta = 3$	0.080	0.080	0.080	0.080	1.000	1.000
SAC-Copula	1.000	1.000	1.000	1.000	0.889	0.919

Table 12: Display-normalized scores used in the v3.8 quality radar. A small display floor is applied to low-scoring axes to avoid degenerate polygons in the visualization. These scores are descriptive display values, not additional primary quality metrics.

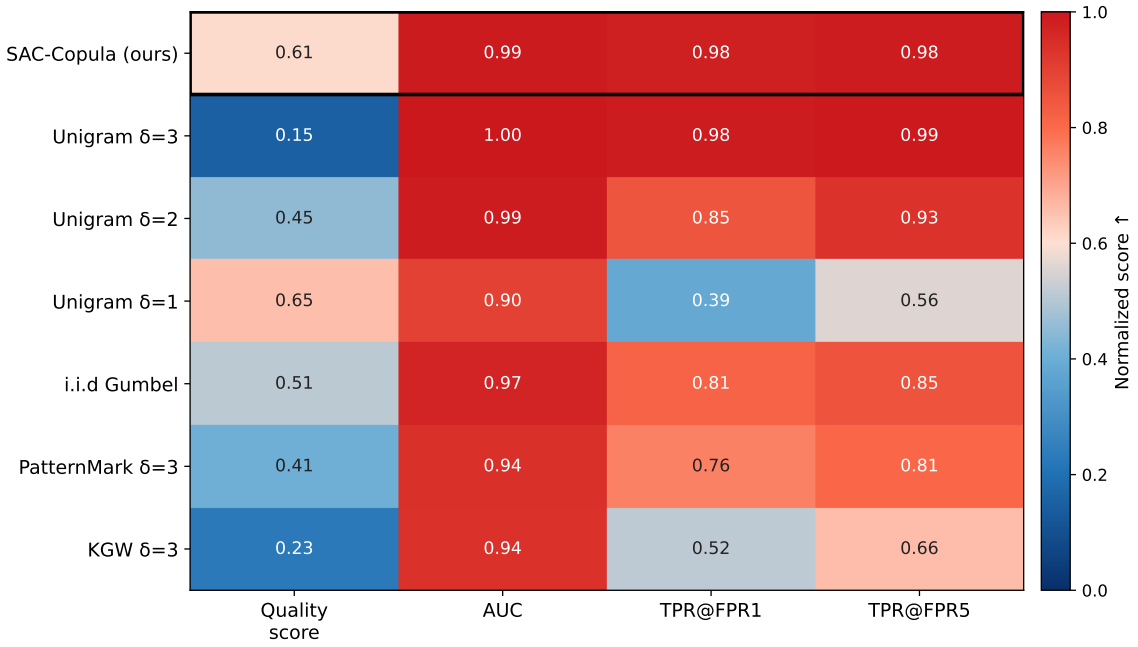


Figure 7: Quality–detection heatmap for the clean multibase comparison. Rows correspond to representative watermark settings, and columns show the normalized quality-preservation score, AUC, TPR@FPR1, and TPR@FPR5. The heatmap quality score averages semantic preservation, MAUVE, token-distribution fidelity, low repetition, lexical diversity, and tail stability after direction-aware normalization. SAC-Copula maintains a favorable balance between clean detectability and quality preservation compared with high-strength external baselines.

System	Output excerpt	
Native	`Eval|Native|701|256`	Dollar stores turn a profit primarily selling items at prices of only a dollar each. This allows them to sell a large quantity of items at a...
i.i.d. Gumbel	`Eval|i.i.d_Baseline|701|256`	Dollar stores turn a profit primarily by selling low-cost items at a lower price point. These items are usually bulk from larger retailers a...
SAC-Copula	`Eval|SAC_W39_S15.0_R0.6|701|256`	Dollar stores turn a profit through selling a wide variety of items at very low prices. They can offer a wide selection of items in from foo...
KGW	`Eval|KGW_ContextGreen_Bias|701|256`	Dollar stores turn a profit by by offering items at discounted prices, which allows them to sell more items per dollar spent. This is done b...
Unigram	`Eval|Unigram_StaticGreen_Bias|701|256`	Dollar stores turn a profit by selling a wide range of goods at low prices. They compete by offering a diverse selection of goods, including...
PatternMark ReferenceLite	`Eval|PatternMark_MarkovPattern_ReferenceLite|701|256`	Dollar stores make a profit through a combination of factors. 1. Lower pricing: Dollar stores sell items at very low prices, often as little...

Table 13: Traceable clean-generation example for the prompt “How do dollar stores turn a profit?” Excerpts are copied from the frozen clean-generation artifact and truncated for space. The table is illustrative only and does not replace the automatic quality metrics.

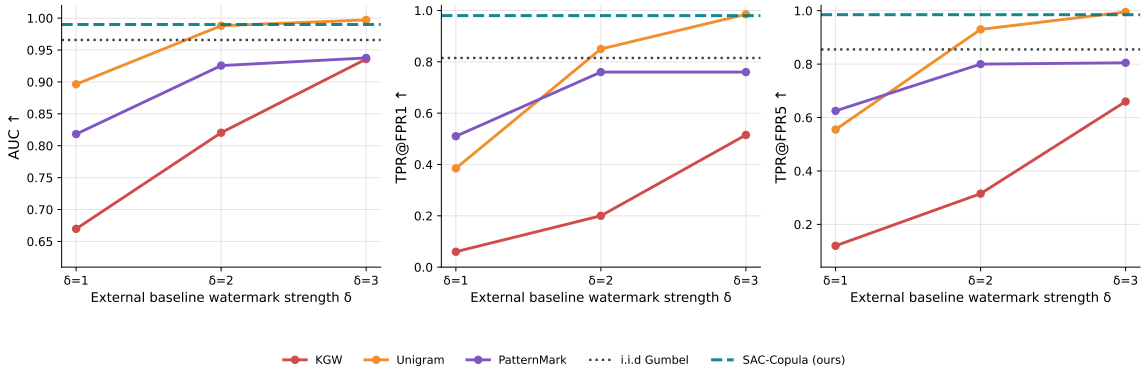


Figure 8: Full clean detection sweep across watermark strengths. The panels report AUC, TPR@FPR1, and TPR@FPR5 for KGW, Unigram, and PatternMark over $\delta \in \{1, 2, 3\}$, with i.i.d. Gumbel and SAC-Copula shown as reference operating points. This figure reports clean fixed-alignment detection only.

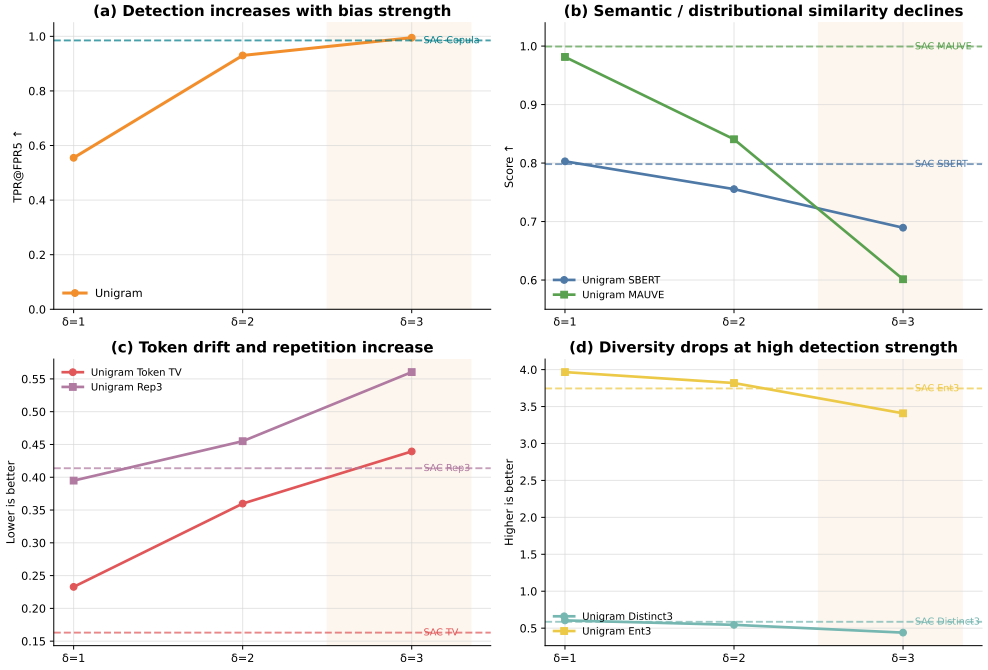


Figure 9: SAC-Copula avoids the high-strength static-bias quality trade-off. The Unigram strength sweep is included as a contrastive reference: Panel (a) shows clean TPR@FPR5 increasing with Unigram strength δ, while Panels (b)–(d) show the accompanying semantic, distributional, token-drift, repetition, and diversity costs. SAC-Copula is shown as the reference operating point.

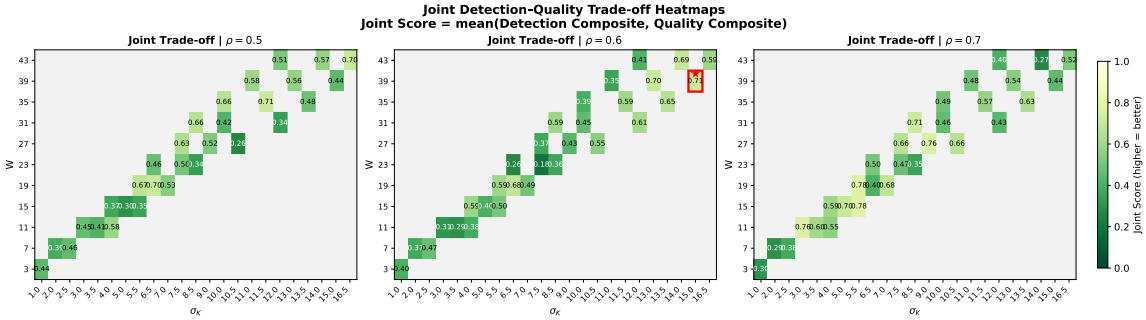


Figure 10: SAC parameter sensitivity for clean operating-point selection. The heatmap groups configurations by ρ, with W and σ_K on the axes and color indicating the joint quality–detectability score. The selected setting $W = 39, \sigma_K = 15.0, \rho = 0.6$ is used as the main SAC-Copula configuration. This visualization guides operating-point selection and is not a claim of global optimality or raw-metric dominance.

Configuration	W	σ_K	ρ	AUC	TPR@1%	TPR@5%	Mean PPL	Collapse	Quality	MAUVE	Rep3	Distinct3	Token TV
SAC W35 S11.5 R0.7	35	11.5	0.7	0.984	0.895	0.965	13.240	0.005	–	0.998	0.443	0.557	0.184
SAC W35 S13.5 R0.7	35	13.5	0.7	0.991	0.950	0.970	11.934	0.000	–	0.980	0.446	0.554	0.189
Selected SAC-Copula	39	15.0	0.6	0.990	0.980	0.985	12.163	0.010	–	0.999	0.414	0.586	0.163

Table 14: Compact appendix summary of SAC parameter sensitivity under clean DLM generation. The selected SAC-Copula row corresponds to the main-text configuration $W = 39, \sigma_K = 15.0, \rho = 0.6$. The normalized composite Quality score is unavailable in the delta-sweep summary CSVs used for this parameter table, so the Quality column is left blank and the table should be interpreted through the listed raw metrics. The $\delta = 1$ data slice is a shared clean-text evaluation slice, not a SAC parameter.

Method / group	Detector	AUC	TPR@1%	TPR@5%	Role
i.i.d. Gumbel	old z-score	0.966	0.815	0.855	Legacy baseline detector.
SAC-Copula	old z-score	0.987	0.915	0.950	Sanity check with the old readout.
SAC-Copula	FFR	0.990	0.980	0.985	Main SAC-aware clean readout.
SAC-Copula	GO-FFR	0.990	0.980	0.985	Offset-enabled variant; same clean score here.
SAC W35 S11.5 R0.7	FFR	0.984	0.895	0.965	Parameter-sweep SAC readout.
SAC W35 S13.5 R0.7	FFR	0.991	0.950	0.970	Parameter-sweep SAC readout.

Table 15: Compact appendix summary of available clean detector comparison rows for i.i.d. Gumbel and SAC-Copula evidence. FFR and GO-FFR read the local correlated evidence generated by SAC, while the old-detector rows provide sanity checks against the legacy i.i.d. readout.

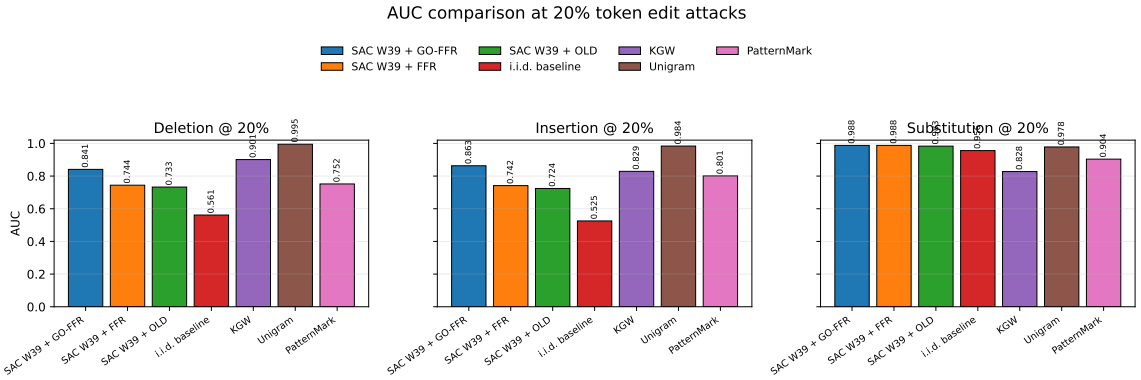


Figure 11: Full AUC comparison under 20% token-level edit attacks. The three panels report deletion, insertion, and substitution for SAC-Copula readouts, the direct i.i.d. Gumbel baseline, and external clean-text baselines. This full comparison provides the complete cross-method view supporting the compact quantitative token-edit discussion in the main text and is reported as a residual-evidence diagnostic, not as a broad robustness claim.

Edit	Detector row	AUC	TPR@1%	TPR@5%	Best offset	Reading
Deletion 20%	i.i.d. old	0.561	0.025	0.110	0.000	Alignment-changing edit strongly weakens i.i.d. evidence.
Deletion 20%	SAC FFR	0.780	0.150	0.370	0.000	SAC-aware filtering retains partial residual evidence.
Deletion 20%	SAC GO-FFR	0.892	0.460	0.590	-27.295	Offset scanning partially recovers coarse shift.
Insertion 20%	i.i.d. old	0.525	0.010	0.045	0.000	Insertions disrupt fixed token-to-tape indexing.
Insertion 20%	SAC FFR	0.842	0.195	0.435	0.000	SAC-aware readout remains informative under controlled insertion.
Insertion 20%	SAC GO-FFR	0.884	0.505	0.670	18.230	Offset scanning gives a coarse synchronization gain.
Substitution 20%	i.i.d. old	0.956	0.755	0.825	0.000	Substitution is less damaging because sequence alignment is preserved.
Substitution 20%	SAC FFR	0.989	0.890	0.975	0.000	SAC evidence remains detectable in the indexed substitution setting.

Table 16: Selected low-rate token-level edit diagnostics from the frozen token-edit summary. These results are reported as partial residual-evidence diagnostics rather than broad robustness claims; the table is not evidence of full edit-path alignment or semantic rewrite robustness.

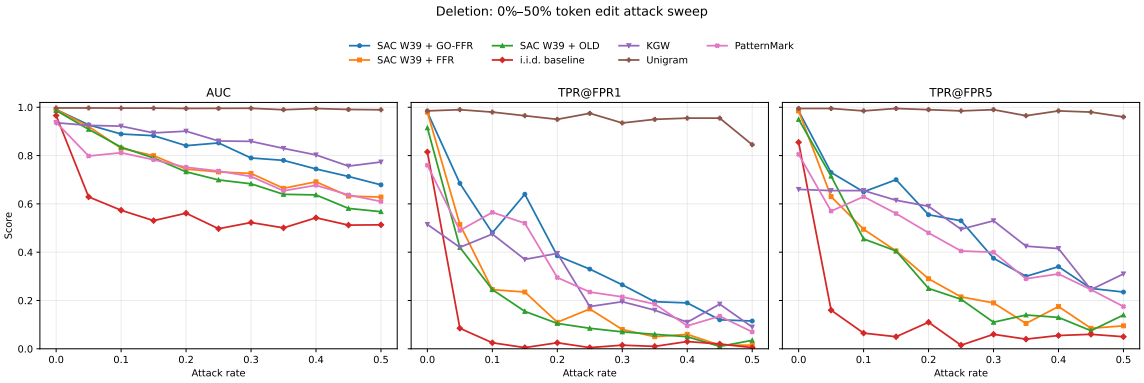


Figure 12: Deletion stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5 for SAC-Copula readouts, i.i.d. Gumbel, and external clean-text baselines. The curves are residual-evidence diagnostics: GO-FFR can partially recover coarse offset shifts, while low-FPR TPR remains limited under accumulated deletion drift.

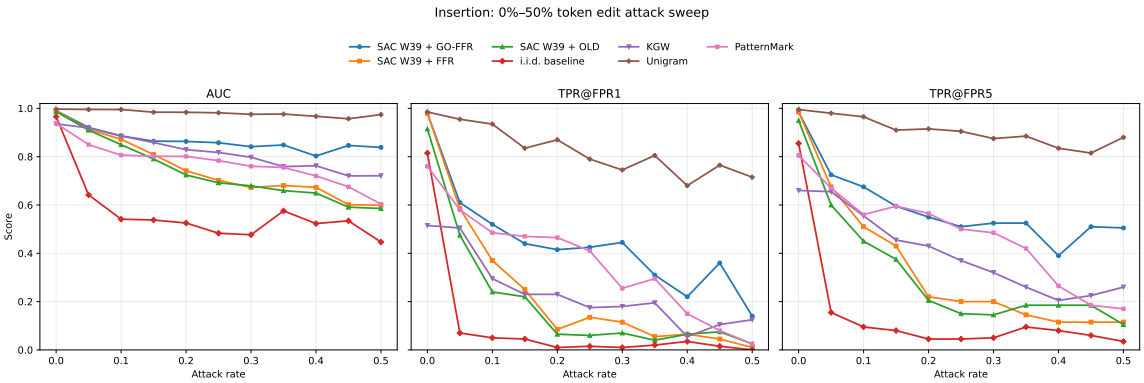


Figure 13: Insertion stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5. Insertion induces synchronization drift by shifting token-to-tape indexing; GO-FFR provides a coarse-offset diagnostic improvement but does not solve cumulative local drift.

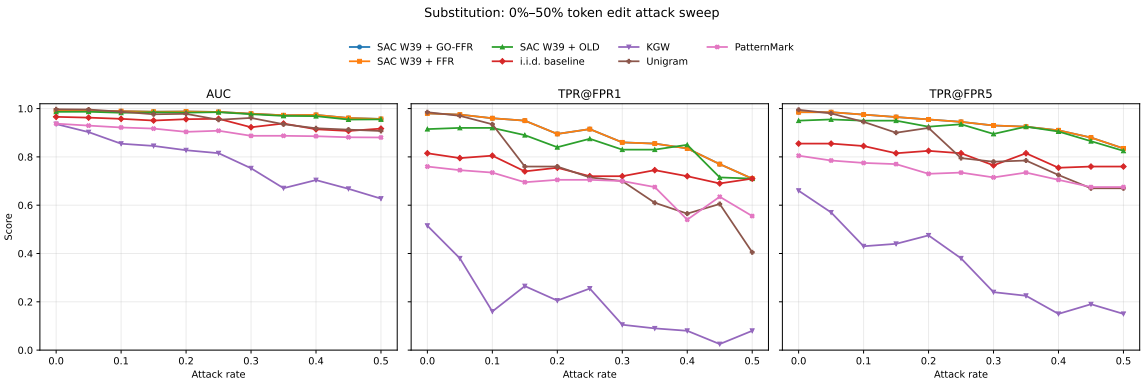


Figure 14: Substitution stress-test sweep from 0% to 50% token edit rate. The three panels report AUC, TPR@FPR1, and TPR@FPR5. Substitution corrupts token identities without shifting all subsequent positions, so it is a controlled token-corruption diagnostic rather than a synchronization-drift setting.

C.8 Failure Analysis under Edit Drift

Edit	Detector	AUC	TPR1	TPR5	Offset
Deletion	i.i.d. old	0.513	0.005	0.050	0.000
	SAC FFR	0.623	0.035	0.085	0.000
	SAC GO-FFR	0.727	0.180	0.285	-51.945
Insertion	i.i.d. old	0.447	0.000	0.035	0.000
	SAC FFR	0.664	0.085	0.160	0.000
	SAC GO-FFR	0.878	0.335	0.550	38.135

Table 17: Diagnostics at 50% deletion/insertion from the frozen token-edit summary. GO-FFR partially improves coarse synchronization, while cumulative local drift remains.

Deletion and insertion accumulate synchronization drift. GO-FFR recovers coarse shifts but not cumulative local drift, leaving edit-path alignment unresolved. Evidence comes from the frozen relative-drop, token-attack, and global-offset summaries.

The frozen semantic/paraphrase summaries cover paraphrase results, valid-only filtering, realized TER bins, relative gains, and detector-only comparisons. They show degraded detection under semantic rewriting; offset-style diagnostics provide only partial recovery.

Attack	Detector	AUC	TPR1	TPR5	Valid
DIPPER R20	i.i.d. row	0.580	0.075	0.105	0.545
DIPPER R20	SAC row	0.745	0.130	0.260	0.495
DIPPER R30	i.i.d. row	0.526	0.005	0.070	0.330
DIPPER R30	SAC row	0.633	0.020	0.140	0.320

Table 18: Semantic/paraphrase failure-analysis rows from the frozen summary, defining the current robustness boundary.

C.9 Tail Stability and Correlation-Strength Controls

The tail analysis in Tables 19–20 re-analyzes 200 frozen LLaDA-8B-Instruct ELI5 outputs for SAC $W = 39, \sigma_K = 15, \rho = 0.6$, paired by `Prompt_ID` and `Gen_Len`. Conditional PPL is evaluated with Meta-Llama-3.1-8B-Instruct; collapse is `Conditional_PPL` > 100. Confidence intervals use 10,000 paired bootstrap resamples with seed 20260712.

Fixed-parameter correlation-strength sweep. The canonical 8B sweep varies $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ while fixing $W = 39$, $\sigma_K = 15$, $\alpha = 1$, temperature 1, generation length 256, block length 32, 128 diffusion steps, and key 42. Prompts, seeds, decoding, and remasking are fixed; both Old and FFR detectors are evaluated, and PPL uses Meta-Llama-3.1-8B-Instruct. The $\rho = 0$ row is the direct matched i.i.d. endpoint. This is a fixed-parameter correlation-strength sweep, not matched-detectability or watermark-strength optimization.

Metric	Native	i.i.d.	SAC
Mean PPL	16.840	636.069	12.163
Median PPL	6.381	8.446	8.260
Q95 trimmed mean	7.201	9.864	9.333
P90	15.387	19.603	17.911
P95	22.580	48.420	30.810
P99	60.653	4256.634	99.653
Maximum	1669.349	59257.688	128.978
PPL>100	1/200	7/200	2/200

Table 19: Conditional-PPL summaries for frozen LLaDA–ELI5 outputs. Q95 is the mean after trimming the highest 5%.

Paired collapse transition	Count
normal $\rightarrow$ normal	192/200
i.i.d. collapse $\rightarrow$ SAC normal	6/200
i.i.d. normal $\rightarrow$ SAC collapse	1/200
collapse $\rightarrow$ collapse	1/200

Diagnostic	Estimate	95% CI
Collapse-rate difference	-0.025	[-0.050, 0.000]
McNemar two-sided	p=0.125	–
Mean $\log_{10}$-PPL difference	-0.0818	[-0.1664,-0.0063]
P99 difference	-4156.981	[-59177.786,-60.720]
Median PPL difference	0.135	[-0.741,1.036]
SAC lower-PPL win rate	0.485	[0.415,0.555]

Table 20: Paired collapse transitions and statistical diagnostics for SAC relative to matched i.i.d. Gumbel. The collapse-rate difference is SAC minus i.i.d.

	AUC	TPR1	TPR5
ρ	Old / FFR	Old / FFR	Old / FFR
0.0	.9658 / .9683	.815 / .820	.855 / .890
0.2	.9671 / .9753	.805 / .865	.895 / .905
0.4	.9819 / .9861	.895 / .925	.950 / .955
0.6	.9866 / .9900	.915 / .980	.950 / .985
0.8	.9842 / .9845	.920 / .925	.950 / .950
1.0	.9748 / .9631	.875 / .830	.940 / .895

Table 21: Complete detection results for the canonical six-point ρ sweep. TPR1 and TPR5 denote TPR at 1% and 5% FPR, respectively.

	Median			PPL	
ρ	PPL	P95	P99	>100	SBERT
0.0	8.45	48.42	4256.63	3.5%	.784
0.2	8.52	36.62	2354.28	2.5%	.792
0.4	8.28	26.00	173.26	2.0%	.797
0.6	8.26	30.81	99.65	1.0%	.798
0.8	8.17	21.97	66.18	0.5%	.795
1.0	7.38	18.00	27.55	0.0%	.766

Table 22: Complete quality-tail results for the canonical six-point ρ sweep. At the two reported endpoints for broader quality context, $\rho = 0.6$ has Rep3 $\approx .414$ and Distinct3 $\approx .586$, while $\rho = 1$ has Rep3 $\approx .528$ and Distinct3 $\approx .472$; unreported intermediate repetition/diversity values are not imputed.

Detection improves toward moderate correlation and then declines, while the quality dimensions do not move uniformly. The sweep therefore supports $\rho = 0.6$ as a balanced interior operating point, not

a universal or global optimum, and does not isolate correlation as a causal theorem.

C.10 Calibration and Threshold-Transfer Controls

Calibration-size sensitivity. These controls use LLaDA–ELI5, SAC $W = 39, \sigma_K = 15, \rho = 0.6$, and model-native H_0. Calibration sizes are $50/100/200/500$, with $20/20/20/1$ repeats, respectively. Development and evaluation identities are fixed and no evaluation example is used for calibration.

N	AUC	TPR@1%FPR	TPR@5%FPR
50	$.9758 \pm .0111$	$.8025 \pm .0715$	$.8982 \pm .0677$
100	$.9867 \pm .0034$	$.9078 \pm .0473$	$.9667 \pm .0184$
200	$.9888 \pm .0009$	$.9487 \pm .0277$	$.9812 \pm .0060$
500	.9900	.9800	.9850

Table 23: ROC discrimination as a function of model-native H_0 calibration size. Mean and standard deviation are reported when repeated subsamples are available.

N	Realized FPR @nom1	Realized FPR @nom5	TPR @nom1	TPR @nom5
50	28.88%	34.33%	98.40%	98.65%
100	13.03%	19.55%	98.08%	98.55%
200	5.13%	11.15%	97.90%	98.53%
500	0.50%	8.50%	98.00%	98.50%

Table 24: Operational-threshold diagnostic across calibration sizes. Strong ROC discrimination does not imply insensitivity of fixed low-FPR thresholds to H_0 sample size.

Calibration-source controls. All four source conditions use $N = 500$, the same development identities, the same 200 LLaDA–ELI5 Native negative evaluations, and the same 200 LLaDA–ELI5 SAC positive evaluations; only the H_0 source changes.

H_0 source	AUC	TPR@1%FPR	TPR@5%FPR
LLaDA–ELI5 Native	.9900	.980	.985
ELI5 human answer	.9878	.955	.980
C4 raw text	.9871	.955	.975
Wikipedia raw text	.9861	.950	.970

Table 25: Four-source H_0 calibration control with fixed development and evaluation identities. The results quantify source sensitivity; no formal equivalence margin was pre-specified.

Threshold portability and prompt shift. Using Wikipedia H_0 thresholds on ELI5 Native evaluation gives realized FPR 6.5% $(13/200)$ and TPR 97.5% at the nominal 1% point, and realized FPR 21.0% $(42/200)$ and TPR 98.5% at the nominal 5% point. Thus, strong ROC separation is not equivalent to threshold portability.

Calibration H_0	Negative Eval	AUC	TPR1	TPR5
Native	Native	.9900	.9800	.9850
Native	ELI5 human	.9908	.9850	.9850
ELI5 human	Native	.9878	.9550	.9800
ELI5 human	ELI5 human	.9902	.9850	.9900

Table 26: Native/Human 2×2 ROC control. This matrix reports score discrimination rather than realized FPR at frozen operational thresholds.

Condition	H_0 construction	AUC	TPR1	TPR5
A	Original 500 ELI5 questions + original starting prompt	.9900	.980	.985
B	Disjoint 500 ELI5 questions + same starting prompt	.9889	.970	.985
C	Same new questions/seeds + alternative starting prompt	.9907	.980	.985

Table 27: Same-model prompt-shift control. A→B changes question/content samples; B→C changes the prompt template with questions and seeds fixed; A→C is the combined tested same-model/same-task shift. This is not a prompt-invariance claim beyond the evaluated setting.

Frozen-threshold C4 `realnewslike` FPR. This detector-only transfer test freezes the original LLaDA–ELI5 model-native-H_0 thresholds, performs no C4 refit, and evaluates $2{,}000$ held-out non-watermarked C4 `realnewslike` records. It is not C4 watermark generation and is distinct from the C4-en continuation experiments below.

Nominal FPR	Realized FPR	Exact 95% CI	SAC TPR
1%	0.95%	[0.57%, 1.48%]	.980
5%	5.50%	[4.54%, 6.59%]	.985

Table 28: Frozen-threshold false-positive transfer from LLaDA–ELI5 Native H_0 to held-out C4 `realnewslike` negatives.

As a secondary diagnostic, rebuilding H_0 on C4 `realnewslike` yields realized FPR 2.95%, exact 95% CI $[2.25\%, 3.79\%]$, at nominal 1%, and 7.55%, CI $[6.43\%, 8.80\%]$, at nominal 5%. This evaluated sample does not show improved false-positive control after rebuilding H_0 on C4; it does not establish that target-domain recalibration is generally harmful. Frozen-threshold realized FPR and ROC-selected detection answer different operational questions.

C.11 Targeted Backbone and Task Transfer

Dream-v0-Instruct-7B on ELI5. This targeted second-backbone evaluation uses 500 held-out

ELI5 human answers for calibration, a separate 200 human answers for negative evaluation, and 200 method-specific Dream generations for positive evaluation. The i.i.d. system uses its original exact-tape detector, while SAC uses FFR.

Detection			
Method	AUC	TPR1	TPR5
i.i.d. Gumbel	1.000000	1.000	1.000
SAC-Copula	.998625	.995	.995

Quality				
Source	Median	P95	P99	Composite collapse
ELI5 human answers	13.42	24.04	30.55	0.0%
i.i.d. Gumbel	30.02	889.58	5988.25	24.5%
SAC-Copula	17.84	81.69	240.68	4.0%

Table 29: Dream-7B detection and quality evidence on ELI5. Composite collapse is the Dream experiment's preregistered PPL/repetition rule and is not the LLaDA `Conditional_PPL` > 100 statistic.

Both methods retain near-perfect detection in this protocol, while SAC reduces the i.i.d. upper-tail and composite-collapse profile. This provides targeted evidence of cross-DLM generalization to a second backbone.

LLaDA on C4-en continuation. The task uses a continuation instruction with a 128-token passage and generates 256 tokens. SAC uses $W = 39, \sigma_K = 15, \rho = 0.6$ with no C4-specific watermark retuning. Calibration uses 200 C4-prompted unwatermarked LLaDA outputs; evaluation uses 500 raw non-watermarked C4-en records and 200 method-specific watermarked outputs. The method-matched pipelines use Old for i.i.d. and FFR for SAC.

Detection			
Pipeline	AUC	TPR1	TPR5
i.i.d.+Old	.8260	.415	.530
SAC+FFR	.9138	.670	.805

Quality			
Method	Median PPL	P99	PPL>100
i.i.d.	13.71	35740.32	10.5%
SAC	13.73	494.57	3.0%

Table 30: Method-matched LLaDA C4-en continuation results. Detection differences are attributed to the complete i.i.d.+Old and SAC+FFR pipelines, not solely to generation-side SAC.

The two pipelines have nearly identical median PPL, while SAC+FFR has a substantially less severe quality tail and stronger detection under the submitted method-matched protocol. This provides targeted evidence of cross-domain generalization to a second task/source setting.

Separate common-detector control. This is a distinct detector-control protocol and is not merged with the method-matched result above.

Detection control				
Method	Old AUC	FFR AUC	FFR TPR1	FFR TPR5
i.i.d.	.9107	.8627	.130	.315
SAC	.9236	.9221	.785	.825

Quality					
Method	Median	P95	P99	PPL > 100	SBERT
i.i.d.	13.71	680.48	35740.32	10.5%	.570
SAC	13.73	78.61	494.57	3.0%	.601

Table 31: Separate C4 common-detector audit. These detector-control values must not be interpreted as the submitted method-matched C4 result in Table 30.

Additional quality context in this separate audit is Rep3 $0.7298 \rightarrow 0.6954$, Distinct3 $0.2702 \rightarrow 0.3046$, and Ent3 $1.7293 \rightarrow 1.9986$ from i.i.d. to SAC. The C4 raw-text calibration-source control, frozen-threshold C4 `realnewslike` FPR test, C4-en method-matched experiment, and this common-detector audit retain their distinct H_0, evaluation, denominator, and detector identities.